

# Adaptive Large Language Model Reasoning for Autonomous Driving Decision-Making

Haicheng Liao<sup>a,1</sup>, Chengyue Wang<sup>a,1</sup>, Hanlin Kong<sup>b</sup>, Zhenning Li<sup>a,\*</sup>, Zhengbing He<sup>c</sup>, Zhiyong Cui<sup>d</sup>, Guohui Zhang<sup>e</sup> and Chengzhong Xu<sup>a,\*</sup>

<sup>a</sup>State Key Laboratory of Internet of Things for Smart City, University of Macau, Macau SAR, China

<sup>b</sup>College of Computer Science and Technology, Zhejiang University, Hangzhou, China

<sup>c</sup>Senseable City Lab, Massachusetts Institute of Technology, Cambridge MA, United States

<sup>d</sup>School of Transportation Science and Engineering, Beihang University, Beijing, China

<sup>e</sup>Department of Civil and Environmental Engineering, University of Hawaii, Honolulu HI, United States

## ARTICLE INFO

### Keywords:

Autonomous Driving  
Trajectory Prediction  
Trajectory Planning  
Large Language Models  
Free Energy Principle

## ABSTRACT

Trajectory prediction and planning are critical for the safe operation of autonomous driving (AD) systems. Existing data-driven and imitation learning pipelines typically rely on geometric pattern fitting and short-horizon statistical correlations, struggling to generalize in complex, interactive scenarios. Although large language models (LLMs) can offer profound semantic understanding and causal reasoning, their massive computational overhead and mismatch with continuous trajectory generation severely hinder their development in real-time AD. In this paper, we propose an LLM-enhanced framework for joint trajectory prediction and planning that combines fast kinematic inference with adaptive semantic reasoning. Inspired by the free energy principle, we introduce a novel Hierarchical Contextual Encoder with an uncertainty-driven trigger. This module dynamically allocates language-model reasoning capacity based on scene complexity and uncertainty, utilizing a lightweight language model (LM) for routine scenarios, and selectively invoking a high-capacity LLM only when deep causal reasoning is required. To enhance the reasoning ability of the LM, we develop a CoT-based latent distillation paradigm that transfers structural reasoning and contextual priors from a teacher LLM to a student model through hidden-space alignment. In addition, we design a Cross-modal decoder that injects semantic intent into geometric trajectory generation, enabling more socially compliant and context-aware decision making. Extensive experiments on multiple real-world benchmarks demonstrate that LLM-TP<sup>2</sup> achieves state-of-the-art performance in both prediction and closed-loop planning, while ensuring the computational efficiency required for practical deployment.

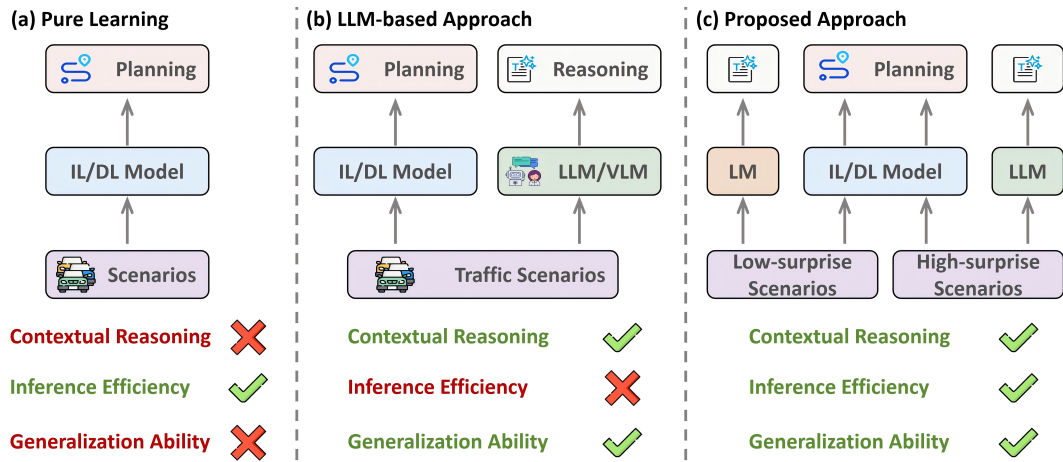
## 1. Introduction

The realization of fully autonomous driving (AD) increasingly necessitates the autonomous vehicle (AV) to jointly anticipate the future movement of surrounding agents (vehicles, pedestrians, and cyclists) and generate safe, feasible, and socially compliant decisions [1]. These two capabilities, trajectory prediction and motion planning, are inherently coupled in real-world traffic, where the future evolution of surrounding agents depends not only on their motion history but also on the intended behavior of the AV itself [2]. While early modular pipelines treated trajectory prediction and planning as decoupled components, modern research has shifted toward unified prediction-and-planning frameworks to mitigate error accumulation and better capture the closed-loop feedback loops inherent in multi-agent interactions.

Despite recent progress, existing joint frameworks still face substantial challenges in real-world deployment. Safe driving is rarely governed by local kinematic cues alone; rather, it demands a deep understanding of contextual semantics, including implicit social norms, long-horizon causal dependencies, and evolving scene dynamics. Conventional rule-based systems [3, 4] lack the flexibility to generalize across unpredictable corner cases, often resulting in “frozen” or overly conservative behavior. As shown in Fig. 1, purely data-driven models [5–7] typically simplify AD decision-making into a sequential forecasting problem. By leveraging deep neural networks to fit historical trajectories, these approaches aim to extract statistical regularities from massive datasets. Although these models can be optimized end-to-end and perform well in common scenarios, they are still largely driven by geometric pattern matching and short-horizon temporal fitting. As a result, they often generalize poorly to unfamiliar or unseen scenes, such as unexpected

\*Corresponding author; <sup>1</sup>Equally Contributed

 zhenningli@um.edu.mo (Z. Li); czxu@um.edu.mo (C. Xu)



**Figure 1:** Comparison of three decision-making paradigms for autonomous driving. Pure learning-based methods are efficient but limited in contextual reasoning, while direct LLM-based approaches enhance reasoning and generalization at the cost of real-time efficiency. The proposed approach addresses this trade-off by adaptively allocating reasoning capacity according to scenario surprise, using a lightweight LM for routine cases and invoking an LLM only when necessary.

human behavior or atypical road conditions, where robust decision-making requires a deeper semantic understanding of evolving multi-agent interactions. To bridge this gap, imitation learning (IL) has emerged as a prominent paradigm for replicating human-like driving expertise. Unlike simple trajectory fitting, IL-based methods [8–10] focus on mimicking expert decision-making policies, seeking to map complex environmental states directly to professional driving actions. However, they still suffer from two fundamental limitations. First, they are vulnerable to distribution shift: when transitioning from open-loop training to closed-loop execution, even small prediction errors can accumulate over time and eventually lead to unsafe behavior. Second, IL models often learn superficial correlations rather than the underlying causes of expert actions. For instance, a planner might learn to decelerate simply because it observes a neighbor slowing down (a learned correlation), without correctly inferring whether the true cause is a red light, an obstacle, or a pedestrian crossing. This raises a fundamental question: *How can we design a framework that inherits the scalability of imitation learning while possessing the causal reasoning depth required for complex driving scenarios?*

Recent advances in large language models (LLMs), such as Qwen [11] and LLaMA-3 [12], offer a potential frontier for enhancing decision-making [13, 14]. With profound semantic understanding and reasoning capabilities, LLMs can provide high-level priors for interpreting complex scenes, inferring latent agent intentions, and facilitating context-aware planning. Nevertheless, directly incorporating LLMs into real-time planning loops remains challenging in practice. A key difficulty is that existing LLM-based approaches often rely on textual descriptions of driving scenes, which are not well-suited for representing fine-grained geometric structure and temporal dynamics. As a result, high-level semantic reasoning does not naturally translate into physically feasible and interaction-safe trajectories. In addition, trajectory generation is inherently a continuous control problem, whereas LLMs are primarily designed for discrete token prediction [15]. This mismatch reduces efficiency, introduces unnecessary computational overhead, and makes large-scale models difficult to deploy within the strict latency constraints of real-world AD systems [16, 17].

Human driving provides an important clue for balancing reasoning depth and computational efficiency. Due to limited cognitive and metabolic resources, they cannot devote the same level of attention and deliberation to every driving scene [18]. In routine scenes, such as highway cruising or car-following, decision-making is largely supported by habitual sensorimotor priors and fast pattern matching. In more complex scenes, such as dense interactions, abrupt motion changes, or limited maneuvering space, drivers tend to allocate more attention and rely on more deliberate reasoning. This behavior is broadly consistent with cognitive theories of surprise and with the free energy principle, which views intelligent agents as adapting inference according to the mismatch between observation and prior expectation [19]. Inspired by this, we argue that LMs should not be integrated into AD pipelines as always-on reasoning engines. Instead, they should be invoked selectively, strictly when the complexity of the current scene exceeds the explanatory capacity of a lightweight model. Guided by this intuition, we introduce a hierarchical reasoning framework



that monitors environmental uncertainty and adaptively activates stronger semantic reasoning only in ambiguous or safety-critical situations, while relying on lightweight models in routine scenes to preserve efficiency.

In this paper, we propose LLM-TP<sup>2</sup>, a hybrid multimodal framework that integrates fast kinematic inference with adaptive semantic reasoning to enable joint trajectory prediction and planning for AVs. The central idea is to combine the efficiency and scalability of lightweight geometric modeling with the contextual reasoning ability of LLMs, while avoiding the main limitations of each. Technically, we rethink the integration of LLMs in AD: rather than tasking them with low-level trajectory generation, we cast them as high-level reasoning modules that guide a real-time prediction-and-planning backbone. To this end, we introduce a Hierarchical Contextual Encoder (HCE) that dynamically allocates reasoning capacity based on the semantic surprise (i.e., epistemic uncertainty) of the current scene. Rather than applying a uniform inference strategy, the HCE modulates computational intensity according to this quantifiable metric. In routine scenes, where environmental dynamics are adequately explained by habitual priors (low surprise), the model relies on a lightweight language model (LM) for rapid semantic reasoning. Conversely, when the observed scene exhibits elevated surprise due to semantically ambiguous or safety-critical interactions, a high-capacity LLM is selectively triggered to perform deeper causal reasoning. This decoupled architecture allows the system to execute complex risk assessments in edge cases without tethering the high-frequency planning cycle to the inference bottlenecks of massive LLMs. To enhance the reasoning capability of the lightweight LM, we introduce a latent Chain-of-Thought (CoT)-based latent distillation paradigm. By aligning the student LM's hidden states directly with the intermediate representations of a teacher LLM, we transfer structured reasoning priors into the compact model. This approach provides the downstream planner with necessary semantic guidance while strictly bounding inference latency. Furthermore, we design a cross-modal attention mechanism to bridge the gap between these semantic features and vectorized scene representations. It projects the latent intents into the geometric coordinate space, forming intention-aware motion codes. In a nutshell, the main contributions of this study are summarized as follows:

- We propose LLM-TP<sup>2</sup>, a unified LLM-enhanced framework for joint prediction and planning that combines fast geometric inference with adaptive semantic reasoning. Inspired by the free energy principle, we introduce a Hierarchical Contextual Encoder equipped with an uncertainty-driven trigger to selectively activate LMs of different capacities based on scene complexity and uncertainty, balancing reasoning capability and efficiency.
- We develop a CoT-based distillation paradigm that transfers structural reasoning and contextual priors from a teacher LLM to a lightweight student LM, enabling efficient semantic reasoning without incurring the full cost of large-model inference. On top of this, we design a Cross-modal Decoder that injects abstract semantic intent into geometric trajectory generation, improving the consistency between scene understanding and motion generation.
- We conduct extensive experiments on various benchmarks to evaluate the proposed framework across semantic understanding, trajectory prediction, and closed-loop planning. The results demonstrate that LLM-TP<sup>2</sup> achieves SOTA performance in safety and robustness, while showcasing its potential for real-world AD deployment.

Overall, the rest of this paper is organized as follows: Section 2 reviews the related literature. Then, Section 3 introduces the problem formulation and details the proposed framework for joint trajectory prediction and planning. Section 4 presents experimental evaluations on multiple real-world datasets. Finally, Section 5 concludes this study.

## 2. Related Work

### 2.1. Integrated Prediction and Planning in Autonomous Driving

Safe autonomous driving requires accurately anticipating the future motions of surrounding agents and effectively leveraging these predictions for decision-making. Early autonomous driving pipelines treated trajectory prediction and planning as independent modules [20–23]. However, this decoupled design suffers from two major drawbacks: (1) errors may propagate and accumulate across stages [24, 25], and (2) it fails to capture the closed-loop coupling between prediction and planning, where the ego vehicle's actions may influence other agents' future behaviors [26–28]. To address these issues, researchers have increasingly explored integrated prediction and planning frameworks. Representative work like MATS [29] proposes an efficient trajectory representation to improve computational scalability when combined with multimodal planning on large-scale AD datasets. To explicitly capture multi-agent interactions, GameFormer [30] incorporates hierarchical game-theoretic reasoning into a Transformer architecture to model interactive behaviors. Furthermore, DiffStack [31] formulates a differentiable modular system for joint optimization

of prediction, planning, and control, while RasterModel [10] unifies prediction and planning by learning planning cost functions directly from data. More recent methods further improve behavioral consistency and representation alignment. For instance, PlanTF [8] introduces a topology-aware formulation for socially compliant multi-agent interactions, whereas HPP [32] and SA-TP<sup>2</sup> [2] propose a hybrid-prediction integrated planning framework that jointly leverages and explicitly aligns occupancy-based and trajectory-based predictions via transformer-based coupling and game-theoretic reasoning, leading to improved prediction-planning consistency and closed-loop planning safety.

Despite these advances, existing integrated prediction and planning methods still struggle to fully understand and reason about complex contexts driving scenarios. In real-world traffic, safe decision making often depends not only on local geometric cues, but also on high-level contextual dependencies such as implicit traffic rules, social conventions, and long-horizon intent evolution. Such contextual reasoning remains challenging for conventional learning-based models, motivating the incorporation of LLMs to enhance scene understanding and reasoning.

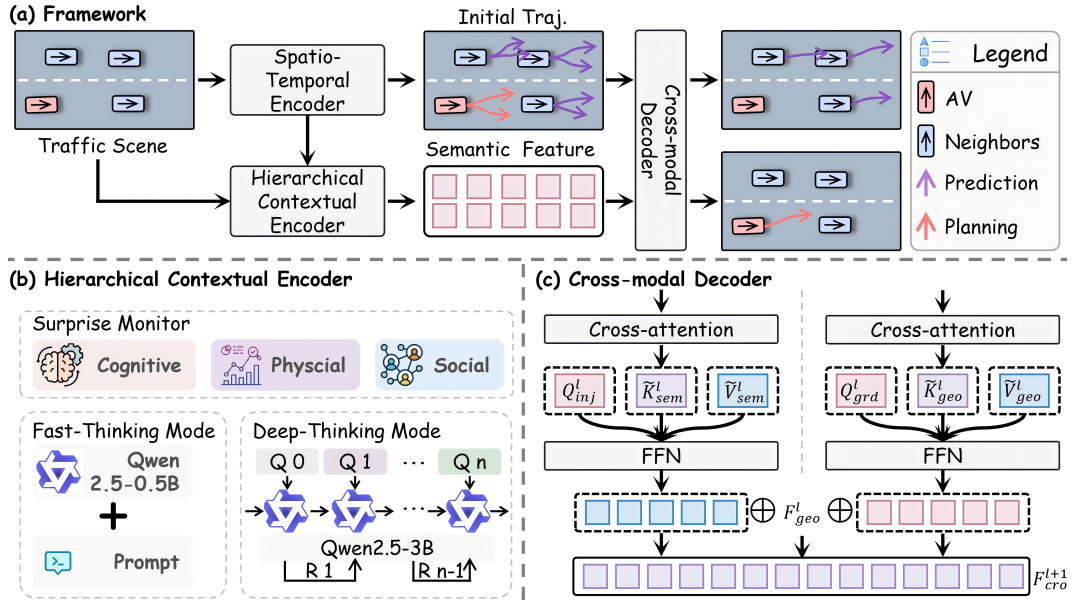
## 2.2. LLMs in Autonomous Driving

Despite the remarkable progress of deep learning in autonomous driving, existing approaches still face challenges in capturing rich contextual semantics and dependencies during rare events and highly interactive scenarios. Recent advances in LLMs have created new opportunities for autonomous driving by improving scene understanding and high-level reasoning [33–35]. In perception and scene understanding, LLMs and vision-language models (VLMs) show strong zero-shot and open-vocabulary capabilities [36]. Prior works [37, 38] show that VLMs can align heterogeneous sensory inputs into unified semantic representations, which are useful for modeling high-level concepts such as traffic rules, social conventions, and latent agent intent. Beyond perception, LLMs are increasingly utilized for interaction modeling and driving decision-making. On the prediction side, frameworks like Traj-LLM [13] and WM-MoE [14] demonstrate that language-guided inference strongly complements standard sequence modeling. Notably, recent studies reframe trajectory prediction as a sentence generation or Visual Question Answering (VQA) task. LMTraj [15] pioneered this by casting trajectory prediction as a sentence generation problem, proving the efficacy of language priors in numerical regression. Building upon this, VLMTraj [39] evolves the task into a comprehensive VQA framework. By converting coordinate sequences and scene images into structured textual prompts and integrating auxiliary multi-task training for group identification and collision assessment, it explicitly guides the model to comprehend complex social relationships and generate socially compliant trajectories. Subsequent studies further leverage VLLMs and question-answering paradigms to enhance prediction flexibility and interpretability across diverse scenarios [15, 40]. On the planning side, LLMs provide a natural interface for representing high-level goals and commonsense constraints, facilitating interpretable behavior planning [41]. To further boost reasoning capabilities, CoT prompting [17] is widely adopted to encourage step-by-step inference, significantly improving robustness [42, 43]. Nevertheless, these LLM-based approaches often suffer from substantial computational overhead, inference latency, and privacy concerns when relying on online services. These practical constraints severely limit their deployment on resource-constrained onboard platforms. Consequently, latency-aware frameworks are needed to preserve the contextual reasoning benefits of LLMs while enabling efficient integration into real-time prediction and planning systems.

## 3. Methodology

### 3.1. Problem Formulation

We formalize the task as a joint problem of probabilistic trajectory prediction for surrounding agents and motion planning for the AV. Consider a traffic scene at the current time  $t$ , comprising the ego vehicle (AV, indexed by 0) and a set of  $n$  surrounding heterogeneous agents (indexed by  $1, \dots, n$ ), such as vehicles, pedestrians, and cyclists. The system observes the environment over a historical horizon  $T_h$ . The input state space is defined as  $\mathcal{I} = \{\mathbf{X}_{0:n}^{t-T_h:t}, \mathbf{M}\}$ . Specifically,  $\mathbf{X}_{0:n}^{t-T_h:t}$  denotes the historical states for the AV and surrounding agents, encompassing kinematic attributes (2D position, velocity, acceleration, yaw angle) and static geometric attributes (length, width). Furthermore,  $\mathbf{M}$  represents the static environmental constraints, including lane graph topology, traffic signal states, and non-traversable static obstacles (curbs, barriers) that define the drivable area boundaries. Given the observation input, the primary objectives are twofold: (i) Trajectory Prediction: To estimate the posterior distribution of future trajectories for all surrounding agents  $i \in \{1, \dots, n\}$  over a prediction horizon  $T_f$ , i.e.  $P(\mathbf{Y}_{1:n}^{t+1:t+T_f} \mid \mathbf{X}_{0:n}^{t-T_h:t}, \mathbf{M})$  where  $\mathbf{Y}_i$  denotes the future state sequence of agent  $i$ ; (2) Ego-agent Planning: To generate a set of  $K$  candidate trajectories  $\mathcal{T}_{ego} = \{\tau_k^*\}_{k=1}^K$  and select the optimal trajectory  $\mathbf{Y}_0^{t+1:t+T_f}$  for the AV, ensuring safety, feasibility, and compliance with traffic rules.



**Figure 2:** Overview of the proposed LLM-TP<sup>2</sup> framework for joint trajectory prediction and planning. (a) The Spatio-Temporal Encoder extracts geometric features and generates initial trajectories, the Hierarchical Contextual Encoder performs adaptive semantic reasoning to produce compact semantic features, and the Cross-modal Decoder fuses both modalities to refine surrounding-agent prediction and ego planning in a coarse-to-fine manner. (b) The Hierarchical Contextual Encoder adaptively switches between Fast-Thinking Mode and Deep-Thinking Mode according to scene surprise. (c) The Cross-modal Decoder integrates features through Cross-modal attention to progressively refine the joint outputs.

### 3.2. Architecture Overview

Figure 2 illustrates the proposed LLM-TP<sup>2</sup> framework, an encoder-decoder architecture inspired by the free energy principle (FEP) for joint trajectory prediction and planning. The framework consists of three main components: (i) a Spatio-Temporal Encoder for modeling geometric scene dynamics, (ii) a Hierarchical Contextual Encoder for semantic reasoning and intent abstraction, and (iii) a Cross-modal Decoder for integrating semantic and geometric information.

The pipeline begins with the Spatio-Temporal Encoder, which models multi-agent interactions and scene dynamics from historical observations. Given the past states of all agents  $X_{0:n}^{t-T_h:t}$  and map context  $M$ , it extracts spatio-temporal geometric features  $F_{geo}$ , predicts the future trajectory distribution of surrounding agents  $P(Y_{1:n}^{t+1:t+T_f})$ , and generates an initial ego proposal set  $\mathcal{T}_{ego}$ . This module serves as a high-frequency geometric prior for downstream reasoning and planning. Next, the Hierarchical Contextual Encoder then performs high-level scene interpretation based on the state space  $I$ , the predicted trajectory distribution  $P(Y_{1:n}^{t+1:t+T_f})$ , and the ego proposals  $\mathcal{T}_{ego}$ . Driven by a surprise-based mechanism under the free energy principle, it adaptively selects either a lightweight LM or a stronger LLM according to the estimated scene complexity. Instead of producing explicit text, the encoder outputs compact semantic latent features  $F_{sem}$ , which encode scene semantics, inferred agent intent, and high-level motion constraints. Finally, the Cross-modal Decoder connects semantic reasoning with geometric trajectory generation. It integrates the spatio-temporal geometric features  $F_{geo}$  with the semantic latent features  $F_{sem}$  through Cross-modal attention, allowing high-level scene semantics to modulate low-level motion generation. It then produces the final joint outputs, including the predicted trajectories of surrounding agents  $Y_{1:n}^{t+1:t+T_f}$  and the final planning trajectory  $Y_0^{t+1:t+T_f}$  for the autonomous vehicle.

### 3.3. Spatio-Temporal Encoder

Serving as the geometric backbone of LLM-TP<sup>2</sup>, the Spatio-Temporal Encoder maps raw historical observations  $I$  into spatio-temporal geometric features  $F_{geo}$ , thereby establishing a robust kinematic prior for future motions. As a high-frequency, data-driven posterior module, it is designed to capture both local agent dynamics and social interactions in a computationally efficient manner. To this end, the encoder is factorized into two complementary components:

spatial interaction modeling (SIM) and temporal dynamics modeling (TDM). They work in synergy to distill spatio-temporal representations, which are processed by a probabilistic decoding head to generate trajectory hypotheses.

### 3.3.1. Vectorized Input Representation

Given the input state space  $\mathcal{I}$ , we first map all heterogeneous elements into a unified latent space  $\mathbb{R}^D$ .

**Dynamic Agent Embedding.** Let  $\mathbf{x}_i^\tau \in \mathbb{R}^{d_{\text{dyn}}}$  denote the raw motion state of agent  $i \in \{0, \dots, n\}$  at time step  $\tau \in \{t - T_h, \dots, t\}$ , including relative position, velocity, and heading information. We use a shared multi-layer perceptron  $\phi_{\text{dyn}}$  to project these kinematic states into a latent feature space:

$$\mathbf{h}_i^\tau = \phi_{\text{dyn}}(\mathbf{x}_i^\tau) \in \mathbb{R}^D, \quad \forall i \in \{0, \dots, n\}, \forall \tau \in [t - T_h, t]. \quad (1)$$

**Static Map Embedding.** We represent the map  $\mathbf{M}$  as a set of  $L$  vectorized polylines, such as lane centerlines and road boundaries. Each polyline  $l$  is discretized into  $G$  segments. Let  $\mathbf{m}_{l,g} \in \mathbb{R}^{d_{\text{map}}}$  denote the  $g$ -th segment of the  $l$ -th polyline, which contains both geometric attributes and semantic attributes, such as road type or traffic constraints. A map-specific MLP  $\phi_{\text{map}}$  projects these static elements into the same latent space:

$$\mathbf{h}_{l,g}^{\text{map}} = \phi_{\text{map}}(\mathbf{m}_{l,g}) \in \mathbb{R}^D, \quad \forall l \in \{1, \dots, L\}, \forall g \in \{1, \dots, G\}. \quad (2)$$

By projecting both dynamic and static inputs to a common dimension  $D$ , the encoder forms a unified token set  $\mathcal{E} = \{\mathbf{h}_i^\tau\} \cup \{\mathbf{h}_{l,g}^{\text{map}}\}$ . This unified representation preserves the heterogeneous structure of the traffic scene and provides the input for subsequent spatio-temporal interaction modeling.

### 3.3.2. Spatial Interaction Modeling

To capture inter-agent dependencies and map constraints, we model spatial interactions using a relational graph attention mechanism. At each time step  $t$ , we construct a heterogeneous interaction graph  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ , where the node set  $\mathcal{V} = \mathcal{V}_a \cup \mathcal{V}_m$  consists of agent nodes  $\mathcal{V}_a = \{0, \dots, n\}$  and map nodes  $\mathcal{V}_m$ . For each agent  $i$ , the spatial feature  $\mathbf{z}_i^t$  is obtained by aggregating messages from its neighborhood  $\mathcal{N}(i) \subseteq \mathcal{V}$ , which can be denoted as  $\mathbf{z}_i^t = \sum_{v \in \mathcal{N}(i)} \alpha_{iv}^t \mathbf{W}_v \mathbf{h}_v^t$ , where  $\mathbf{W}_v$  is a learnable projection depending on the node type. The attention weight  $\alpha_{iv}^t$  is defined as follows:

$$\alpha_{iv}^t = \frac{\exp(\text{score}(\mathbf{h}_i^t, \mathbf{h}_v^t, \mathbf{e}_{iv}))}{\sum_{u \in \mathcal{N}(i)} \exp(\text{score}(\mathbf{h}_i^t, \mathbf{h}_u^t, \mathbf{e}_{iu}))}, \quad (3)$$

where  $\mathbf{e}_{iv}$  encodes explicit geometric and topological relations, such as relative distance, lane offset, and traffic-state information. The scoring function is implemented as  $\text{score}(\mathbf{h}_i^t, \mathbf{h}_v^t, \mathbf{e}_{iv}) = \phi_{\text{gat}}([\mathbf{W}_q \mathbf{h}_i^t \parallel \mathbf{W}_k \mathbf{h}_v^t \parallel \mathbf{e}_{iv}])$ , where  $\parallel$  denotes concatenation, and  $\phi_{\text{gat}}$ ,  $\mathbf{W}_q$ , and  $\mathbf{W}_k$  are learnable mappings. By incorporating relational cues directly into attention, the encoder can focus on interaction-critical agents and map elements, producing scene-aware spatial features that capture both collision-relevant interactions and map-imposed feasibility constraints.

### 3.3.3. Temporal Dynamics Modeling

Given the interaction-aware spatial features at each historical time step, we model temporal evolution using a temporal self-attention mechanism. Specifically, for each agent  $i \in \{0, \dots, n\}$ , the sequence of spatial features  $\{\mathbf{z}_i^\tau\}_{\tau=t-T_h}^t$  is processed by a temporal encoder to capture both short-term motion variations and longer-term kinematic dependencies. Hence, the temporal features can be computed as  $\mathbf{u}_i = \sum_{\tau=t-T_h}^t \beta_{i,\tau} \mathbf{W}_i \mathbf{z}_i^\tau$ , where  $\mathbf{u}_i$  denotes the temporal feature of agent  $i$ ,  $\mathbf{W}_i$  is a learnable projection matrix, and  $\beta_{i,\tau}$  is the attention weight assigned to time step  $\tau$ . Moreover, the temporal attention weights can be formally defined as follows:

$$\beta_{i,\tau} = \frac{\exp((\bar{\mathbf{W}}_q \mathbf{z}_i^t)^\top (\bar{\mathbf{W}}_k \mathbf{z}_i^\tau))}{\sum_{\tau'=t-T_h}^t \exp((\bar{\mathbf{W}}_q \mathbf{z}_i^t)^\top (\bar{\mathbf{W}}_k \mathbf{z}_i^{\tau'}))} \quad (4)$$

where  $\bar{\mathbf{W}}_q$  and  $\bar{\mathbf{W}}_k$  are learnable query and key projections. To preserve temporal ordering, continuous-time positional encodings are added to  $\mathbf{z}_i^\tau$  before attention is applied. Since the spatial features already encode map constraints and interaction structure, the temporal encoder further captures how these interaction patterns evolve over the historical window, producing a temporally enriched representation for downstream prediction and planning.

### 3.3.4. Spatio-Temporal Feature Fusion

After spatial interaction modeling and temporal dynamics encoding, we fuse the two streams into the unified spatio-temporal geometric features. For each agent  $i$ , the current spatial features  $\mathbf{z}_i^t$  and the temporal features  $\mathbf{u}_i$  are synthesized through a gated fusion mechanism, denoted as  $\mathbf{g}_i = \sigma(\mathbf{W}_g [\mathbf{z}_i^t \parallel \mathbf{u}_i])$ ,  $\tilde{\mathbf{h}}_i = \mathbf{g}_i \odot \mathbf{u}_i + (1 - \mathbf{g}_i) \odot \mathbf{z}_i^t$ , where  $[\cdot \parallel \cdot]$  denotes feature concatenation,  $\sigma(\cdot)$  is the sigmoid activation, and  $\odot$  denotes element-wise multiplication. This gating mechanism adaptively balances instantaneous interaction cues and historical motion context for each agent.

To derive a comprehensive scene-level geometric encoding, the fused agent embeddings  $\{\tilde{\mathbf{h}}_i\}_{i=0}^n$  are aggregated using an attention-based pooling operator. This mechanism enables the encoder to selectively emphasize interaction-critical agents while suppressing irrelevant noise. The attention weights  $\alpha_i$  are computed by measuring the relative importance of each agent embedding, which can be represented as  $\alpha_i = \frac{\exp(\mathbf{w}^\top \tilde{\mathbf{h}}_i)}{\sum_{j=0}^n \exp(\mathbf{w}^\top \tilde{\mathbf{h}}_j)}$ . Then, the global spatio-temporal geometric features are formulated as the weighted sum of the individual agent features, denoted as  $F_{\text{geo}} = \sum_{i=0}^n \alpha_i \tilde{\mathbf{h}}_i$ . Here, the geometric features  $F_{\text{geo}}$  provide a compact spatio-temporal summary of the entire traffic environment, the agent-specific features  $\tilde{\mathbf{h}}_i$  are retained to support downstream multi-agent trajectory distribution modeling.

### 3.3.5. Probabilistic Trajectory Decoding

The final stage of the Spatio-Temporal Encoder is to convert the spatio-temporal geometric features  $F_{\text{geo}}$  into explicit motion states. This module provides a kinematic prior over future behaviors before semantic refinement. Specifically, we employ an MLP-based decoding head to regress the parameters of a Gaussian Mixture Model (GMM), thereby capturing the multi-modal nature of future traffic motion.

For each surrounding agent  $i \in \{1, \dots, n\}$ , the module predicts a mixture distribution over future trajectories, capturing the inherent multi-modality of traffic behaviors, such as lane following, turning, acceleration, and deceleration. Specifically, it maps the agent-specific feature vector  $f_{\text{geo},i}$  (the  $i$ -th row of  $F_{\text{geo}}$ ) to a set of GMM parameters  $\Theta_i = \{\pi_{i,k}, \mu_{i,k}, \Sigma_{i,k}\}_{k=1}^K$ , where  $K$  denotes the number of motion modes. These modes account for diverse future behaviors, including different lateral and longitudinal maneuvers such as lane following, turning, acceleration, and deceleration. The resulting predictive distribution over the future trajectory  $\mathbf{Y}_i^{t+1:t+T_f}$  can be formulated as follows:

$$P(\mathbf{Y}_i | F_{\text{geo}}) = \sum_{k=1}^K \pi_{i,k} \cdot \mathcal{N}(\mathbf{Y}_i | \mu_{i,k}, \Sigma_{i,k}) \quad (5)$$

Here,  $\pi_{i,k}$  denotes the mixture weight of the  $k$ -th mode and satisfies  $\sum_k \pi_{i,k} = 1$ . The mean  $\mu_{i,k} \in \mathbb{R}^{T_f \times 2}$  represents the predicted waypoint sequence of the corresponding mode, while  $\Sigma_{i,k}$  models the associated aleatoric uncertainty. For numerical stability, the covariance at each future step is parameterized as a bivariate Gaussian with standard deviations  $(\sigma_x^\tau, \sigma_y^\tau)$  and correlation coefficient  $\rho^\tau$ . In this way, the decoder produces a probabilistic kinematic prior that captures both mode diversity and trajectory uncertainty for downstream prediction and planning.

For the ego agent, the decoder generates a finite set of trajectory candidates that serve as anchors for subsequent semantic evaluation and refinement. Conditioned on the ego-specific geometric feature  $f_{\text{geo},0}$ , the module outputs  $K$  most probable mixture components and uses their mean trajectories as ego-motion proposals. Formally,

$$\mathcal{T}_{\text{ego}} = \{\tau_k^*\}_{k=1}^K, \text{ where } \tau^k = \mu_{0,k}^{t+1:t+T_f} \quad (6)$$

These candidate trajectories  $\tau_k^*$  correspond to the most likely kinematically feasible motion modes predicted by this module. They provide a compact proposal set that preserves the vehicle's local motion tendency, such as lane following or turning preference, while leaving room for subsequent semantic refinement. Next, these proposals are further evaluated and adjusted by the Cross-modal Decoder under high-level semantic guidance.

## 3.4. Hierarchical Contextual Encoder

As shown in Fig. 2 (b), the Hierarchical Contextual Encoder is the semantic reasoning module of LLM-TP<sup>2</sup>, responsible for transforming the input state space  $\mathcal{I}$  into compact semantic latent features  $F_{\text{sem}}$ . Its design is motivated by the resource-rational nature of human cognition: drivers do not reason with the same intensity in every scene, but instead adapt their cognitive effort to the complexity of the environment. In routine situations, such as highway cruising or sparse traffic, decision-making is largely supported by habitual priors and fast pattern matching. In contrast, more



deliberate reasoning is required in complex or unfamiliar scenes, such as dense urban intersections or long-tail corner cases. This adaptive behavior is closely related to the free energy principle, which interprets intelligent inference as minimizing environmental surprise. In our framework, surprise is approximated by variational free energy (VFE), whose magnitude serves as a control signal for reasoning depth. When the VFE is low, the current scene is well explained by the model's internal priors, and the HCE operates in a fast-thinking mode using a lightweight language model. When the VFE is high, indicating elevated epistemic uncertainty or potential safety risk, the encoder switches to a deep-thinking mode that invokes a stronger LLM for richer causal reasoning. In this way, semantic inference is performed adaptively rather than uniformly, allowing the model to preserve efficiency in routine scenes while allocating stronger reasoning capacity to challenging cases. Accordingly, the HCE consists of two components: a Hierarchical Surprise Monitor, which estimates scene uncertainty through multi-dimensional surprise signals, and an Adaptive Reasoning Engine, which selects the appropriate language model and produces the semantic latent feature  $F_{\text{sem}}$ .

### 3.4.1. Hierarchical Surprise Monitor

This module estimates the level of scene surprise and uses it to determine the required reasoning depth of the model. In information theory, surprise is defined as the negative log-likelihood of an observation,  $S(\mathcal{I}) = -\ln p(\mathcal{I})$ . Since the true data distribution is generally intractable in real-world driving, we use VFE as a tractable surrogate. Under the free energy principle, minimizing VFE corresponds to minimizing an upper bound on surprise, thereby encouraging the internal model to remain consistent with the observed environment. To reflect the multi-faceted nature of driving uncertainty, we decompose scene surprise into three complementary dimensions: (1) cognitive surprise, (2) physical surprise, and (3) social surprise. These multi-layered surprises capture both the model's internal uncertainty and the external risks arising from the traffic environment, and provide the basis for adaptive reasoning in the HCE.

**Cognitive Surprise.** Cognitive surprise, denoted by  $S_{\text{cog}}$ , measures the epistemic uncertainty of the Spatio-Temporal Encoder with respect to its own predictive inference. A high value of  $S_{\text{cog}}$  indicates that the current scene  $\mathcal{I}$  cannot be adequately explained by the encoder's learned priors, and that the model is uncertain about the future evolution of the scene. For the  $i$ -th agent, we define  $S_{\text{cog}}(i)$  as the variational predictive entropy of the posterior trajectory distribution  $P(\mathbf{Y}_i | F_{\text{geo}})$ . This uncertainty reflects two aspects of the driving task: ambiguity in maneuver selection and uncertainty in trajectory execution. Since the exact Shannon entropy of a GMM has no closed-form expression, we adopt a second-order Taylor approximation around the component means. Mathematically,

$$S_{\text{cog}}(i) \approx -\sum_{k=1}^K \pi_{i,k} \ln \pi_{i,k} + \frac{1}{2} \sum_{k=1}^K \pi_{i,k} \ln((2\pi e)^T |\Sigma_{i,k}|), \quad (7)$$

where  $\pi_{i,k}$  and  $\Sigma_{i,k}$  denote the mixture weight and covariance matrix of the  $k$ -th mode, respectively, and  $T$  is the prediction horizon. Notably, the first term corresponds to the entropy over maneuver modes and captures uncertainty at the intent level. In scenarios such as multi-agent intersections where the encoder cannot definitively distinguish between conflicting maneuvers, this term increases significantly to signal a semantic intent bifurcation. Correspondingly, the second term, representing the expected differential entropy, evaluates the precision of spatial execution. Larger covariance implies that the predicted trajectory is weakly constrained by the current observations and map structure, resulting in a state of kinematic diffusion. Together, these two terms provide a compact estimate of the model's cognitive uncertainty. When  $S_{\text{cog}}$  exceeds a predefined threshold, the model interprets the current scene as insufficiently explained by the fast kinematic encoder and activates the slow reasoning mode for further semantic disambiguation.

**Physical Surprise.** Physical surprise, denoted by  $S_{\text{phy}}$ , measures the extent to which the proposed trajectories violate physical constraints and safety boundaries. It quantifies the mismatch between the kinematic proposals generated by the encoder and the basic requirements of safety, feasibility, and road compliance. We decompose this surprise into two components: collision surprise, which evaluates spatial safety with respect to the environment and surrounding agents, and kinematic surprise, which evaluates the dynamic legitimacy of the agent motion.

**Collision Surprise.** This metric identifies the spatial risk of an ego-motion proposal by evaluating its proximity to non-drivable regions and surrounding traffic participants. We formulate it as a potential-field-based energy that penalizes trajectories approaching restricted areas or other agents. For the ego-vehicle's proposal  $\tau_k^* \in \mathcal{T}_{\text{ego}}$ , the

collision surprise over the planning horizon  $T_f$  can be formally defined as follows:

$$S_{col}(k) = \sum_{\tau=t+1}^{t+T_f} \left[ \underbrace{D_{sdf}(\mathbf{y}_\tau^k, \mathbf{M})}_{\text{Boundary Violation}} + \gamma \sum_{i=1}^n \underbrace{\Psi(\mathbf{y}_{\tau^k}, \hat{\mathbf{y}}_{i,\tau})}_{\text{Inter-agent Proximity}} \right] \quad (8)$$

where the  $\mathbf{y}_{0,k}^\tau$  denotes the ego waypoint of proposal  $k$  at future time  $\tau \in \{t+1, \dots, t+T_f\}$ , and  $\hat{\mathbf{y}}_{i,\tau}$  denotes the most likely future position of the  $i$ -th surrounding agent. The parameter  $\gamma \in (0, 1)$  is a weighting factor that places greater emphasis on imminent risks.  $D_{sdf}(\cdot)$  is a signed distance penalty that becomes positive when the ego waypoint enters non-drivable regions or static obstacles defined by  $\mathbf{M}$ . Furthermore, the second term utilizes a repulsive Gaussian kernel centered at  $\hat{\mathbf{y}}_{i,\tau}$ , which denotes the most probable future position of agent  $i$ . This kernel serves to penalize the vehicle's proximity to surrounding agents and can be formally defined as follows:

$$\Psi(\mathbf{y}_\tau^k, \hat{\mathbf{y}}_{i,\tau}) = \exp\left(-\frac{\|\mathbf{y}_\tau^k - \hat{\mathbf{y}}_{i,\tau}\|^2}{2\sigma_{safe}^2}\right) \quad (9)$$

where  $\sigma_{safe}$  controls the spatial extent of the repulsive field. A large value of  $S_{col}$  indicates that the current proposal does not provide sufficient spatial clearance with respect to nearby obstacles or traffic participants.

**Kinematic Surprise.** Kinematic surprise  $S_{kin}$  measures whether the proposed motion remains dynamically feasible and behaviorally smooth. It penalizes violations of longitudinal, lateral, and temporal motion limits, thereby identifying trajectories that are physically difficult to execute or overly abrupt for normal driving. For  $\tau_k^* \in \mathcal{T}_{ego}$ , the kinematic surprise  $S_{kin}$  for the  $k$ -th proposal is formally formulated as follows:

$$S_{kin}(k) = \sum_{\tau=t+1}^{t+T_f} (\lambda_a \phi_{ReLU}(|a_\tau| - a_{com}) + \lambda_j \phi_{ReLU}(|j_\tau| - j_{max}) + \lambda_{lat} \phi_{ReLU}(v_\tau^2 |\kappa_\tau| - a_{lat,max})) \quad (10)$$

where  $a_\tau$  and  $j_\tau$  denote the longitudinal acceleration and jerk, respectively, and  $v_\tau^2 |\kappa_\tau|$  corresponds to the lateral centripetal acceleration. To ensure the trajectory is steerable and grounded in physical reality, we constrain the curvature  $\kappa_\tau$  using the Kinematic Bicycle Model. Under this model, the instantaneous curvature is bounded by the mechanical limits of the steering system:  $|\kappa| \leq \frac{\tan \delta_{max}}{L_{veh}}$ , where  $\delta_{max}$  is the maximum steering angle and  $L_{veh}$  is the vehicle wheelbase. The thresholds  $a_{com}$ ,  $j_{max}$ , and  $a_{lat,max}$  are chosen according to vehicle dynamics and comfort limits. A large value of  $S_{kin}$  indicates that the current proposal is dynamically unstable, physically infeasible, or excessively abrupt. Consequently, the total physical surprise  $S_{phy}$  can be formally defined as  $S_{phy} = w_{col} S_{col} + w_{kin} S_{kin}$ , where  $w_{col}$  and  $w_{kin}$  are learnable parameters. A high value of  $S_{phy}$  indicates the kinematic prior is no longer sufficiently consistent with the physical constraints, and thus motivates the ARE to invoke deeper reasoning.

**Social Surprise.** Social surprise, denoted by  $S_{soc}$ , characterizes the degree of social conformity by quantifying the alignment between the ego vehicle's intended motion and the collective behavior manifold of the surrounding agents. Complementing the hard constraints of physical feasibility, social surprise evaluates the alignment of the ego-vehicle with the collective behavioral manifold. This distinction allows the system to extend its safety boundary beyond mere collision avoidance by identifying socially deviant maneuvers, such as aggressive cut-ins, which violate the implicit etiquette of the traffic environment even when they satisfy kinematic limits. To capture the multi-scale nature of traffic interactions, we partition this metric into local conformity surprise, which measures alignment with immediate neighbors, and macroscopic stability surprise, which evaluates the stability of the broader traffic regime.

**Local Conformity Surprise.** The first surprise measures whether an ego-motion proposal is consistent with the local motion patterns of surrounding agents. It models local traffic flow through a non-parametric Kernel Density Estimation (KDE) approach to construct a probabilistic flow field. We define this surprise as the negative log-likelihood of the ego-proposal  $\tau_k^*$  conforming to the directional and kinematic patterns of the neighborhood  $\mathcal{N}_{roi}$ . For each trajectory proposal, the local conformity surprise is formulated as follows:

$$S_{loc}(k) = 1 - \frac{1}{H} \sum_{j \in \mathcal{N}_{roi}} \underbrace{\exp\left(-\frac{1 - \cos(\theta_0^\tau, \theta_j^\tau)}{\sigma_{dir}^2}\right)}_{\text{Directional Alignment}} \cdot \underbrace{\exp\left(-\frac{\|\mathbf{v}_0^\tau - \mathbf{v}_j^\tau\|^2}{\sigma_{vel}^2}\right)}_{\text{Velocity Coherence}} \cdot \underbrace{\exp\left(-\frac{\|\mathbf{p}_0^\tau - \mathbf{p}_j^\tau\|^2}{\sigma_{pos}^2}\right)}_{\text{Spatial Proximity}} \quad (11)$$

where  $\theta_0^\tau, \mathbf{v}_0^\tau, \mathbf{p}_0^\tau$  denote the heading, velocity, and position of the ego proposal at time  $\tau$ , respectively, and the subscript  $j$  refers to neighboring agents in the perception range. The kernel bandwidths  $\sigma_{\text{dir}}$ ,  $\sigma_{\text{vel}}$ , and  $\sigma_{\text{pos}}$  control the sensitivity to heading deviation, velocity difference, and spatial distance, respectively, while  $H$  is a normalization constant. The three kernels jointly encourage alignment in motion direction, speed, and spatial neighborhood.

**Macroscopic Stability Surprise.** Complementing the microscopic interaction perspective, the macroscopic stability surprise  $S_{\text{mac}}$  evaluates the degree of local macroscopic flow instability in the traffic environment surrounding the ego vehicle. This metric is designed to capture traffic regimes in which collective effects, such as congestion wave formation, speed dispersion, and transient flow breakdown, make purely kinematic predictions less reliable. To this end, we formulate the surprise metric as a density modulated instability potential:

$$S_{\text{mac}} = \frac{1}{T_f} \sum_{\tau=t+1}^{t+T_f} \phi_{\text{Sig}} \left( \frac{k_\tau - k_{\text{crit}}}{\sigma_k} \right) \left[ \lambda_{\text{wave}} \phi_{\text{Soft}} \left( -\frac{\tilde{w}_\tau}{\sigma_r} \right) + \lambda_{\text{press}} \phi_{\text{Soft}} \left( \frac{P_\tau - P_{\text{safe}}}{\sigma_p} \right) \right] \quad (12)$$

where  $\phi_{\text{Sig}}(\cdot)$  denotes the Sigmoid activation function;  $\phi_{\text{Soft}}$  denotes the Softplus activation function;  $k_\tau$  is the local traffic density at prediction step  $\tau$ ;  $k_{\text{crit}}$  denotes the critical density above which the traffic flow enters a more congested and potentially unstable regime;  $\tilde{w}_\tau$  is a wave propagation indicator, with negative values corresponding to stronger backward moving congestion wave tendencies;  $P_\tau$  is a pressure like measure characterizing the joint effect of density and velocity dispersion in the local traffic flow;  $P_{\text{safe}}$  denotes the reference threshold for acceptable traffic pressure;  $\sigma_k$ ,  $\sigma_r$ , and  $\sigma_p$  are scaling factors; and  $\lambda_{\text{wave}}$  and  $\lambda_{\text{press}}$  are weighting coefficients.

Following this, we integrate these indicators into a social surprise metric:  $S_{\text{soc}} = w_{\text{loc}} S_{\text{loc}} + w_{\text{mac}} S_{\text{mac}}$ , where  $w_{\text{loc}}$  and  $w_{\text{mac}}$  are learnable parameters. Elevated values signal that the reflexive kinematic prior diverges significantly from the collective consensus, manifesting as either a local interaction conflict or a macroscopic stability risk.

**Surprise Aggregation and Uncertainty-driven Trigger.** To consolidate the multi-dimensional uncertainty into a unified decision-making signal, the individual surprise components are normalized and aggregated into a VFE score:  $S_{\text{total}} = w_{\text{cog}} S_{\text{cog}} + w_{\text{phy}} S_{\text{phy}} + w_{\text{soc}} S_{\text{soc}}$ . In this formulation, the coefficients  $\{w_{\text{cog}}, w_{\text{phy}}, w_{\text{soc}}\}$  represent the strategic importance weights assigned to the cognitive, physical, and social dimensions, respectively. Then, this aggregate score is subsequently normalized to a range of  $[0, 1]$ , denoted as  $\tilde{S}_{\text{total}}$ . This aggregate score functions as an adaptive cognitive controller that dynamically regulates the model's computational routing to maintain an optimal balance between inference latency and reasoning depth. Specifically, the routing decision is governed by a switching threshold  $\psi_{\text{switch}}$ . To explicitly optimize the balance between real-time inference efficiency and deep reasoning capability, we anchor this threshold at a high percentile of the normalized surprise distribution on the validation set. In our design, we calibrate the threshold between the 90th and 95th percentiles of the historical distribution, dynamically adapting it to the specific complexity of diverse benchmarks. Consequently, this data-driven calibration ensures that the highly efficient lightweight LM processes the vast majority of routine maneuvers, strictly reserving the computationally intensive deep-thinking mode for the top 5% to 10% most uncertain interactive scenes. When the total surprise satisfies  $\tilde{S}_{\text{total}} \leq \psi_{\text{switch}}$ , the current scene is regarded as routine and sufficiently explained by the lightweight LM, and the model remains in the fast-thinking mode ( $\delta_{\text{mode}} = 0$ ). Conversely, when  $\tilde{S}_{\text{total}} > \psi_{\text{switch}}$ , the scene is considered too complex for the lightweight model, triggering the deep thinking mode ( $\delta_{\text{mode}} = 1$ ) to invoke the stronger LLM. The binary signal  $\delta_{\text{mode}}$  serves as the primary trigger for the subsequent Adaptive Reasoning Engine. By strictly reserving the computationally intensive LLM for the most challenging interactive minority, this design successfully caps the average computational overhead while maximizing semantic disambiguation in corner cases.

### 3.4.2. Adaptive Reasoning Engine

Guided by the binary routing signal  $\delta_{\text{mode}}$  produced by the HCE, the ARE maps the input state space  $\mathcal{I}$  to the compact semantic latent feature  $F_{\text{sem}}$ , which serves as high-level guidance for downstream trajectory generation. Rather than applying a uniform reasoning strategy to all scenes, the engine adaptively selects between two semantic inference pathways according to the estimated scene complexity. Specifically,  $\delta_{\text{mode}}$  determines whether the model uses a lightweight language model for efficient semantic abstraction or a stronger multimodal LLM for deeper reasoning. In both cases, the generated text is not used directly for prediction or planning. Instead, the internal hidden representations of the selected model are projected into the semantic latent space and passed to the downstream Cross-modal Decoder.

**Deep-Thinking Mode.** When  $\delta_{\text{mode}} = 1$ , the current scene is considered too uncertain or complex to be handled reliably by the lightweight model, and the engine switches to a frozen multimodal LLM, Qwen2.5-VL-3B, for deeper semantic reasoning. To encourage causally grounded scene understanding, we use a structured CoT prompt that guides

the model to focus on four aspects: environmental topology and map constraints, potential collision risks and their causes, the likely intentions and future evolution of interacting agents, and a high-level safety-aware strategy for the ego vehicle. Conditioned on  $\mathcal{I}$ , together with the CoT prompt and task instructions, the model performs semantic reasoning at a higher level of abstraction. As in the fast pathway, we do not use the generated text directly. Instead, the hidden representations of the LLM are projected into the same semantic latent space to obtain  $F_{\text{sem}}$ , which encodes scene semantics, inferred intent, and high-level strategic guidance in a compact continuous representation. Through this dual-pathway design, the ARE transforms the free-energy-based routing decision into semantic features of different reasoning depths, allowing the model to adjust its semantic inference cost according to scene complexity.

**Fast-Thinking Mode.** When  $\delta_{\text{mode}} = 0$ , the current scene is considered sufficiently explained by the lightweight model, and the engine operates in the fast-thinking mode. In this regime, the goal is efficient semantic abstraction rather than detailed deliberation. We therefore use a compact language model, Qwen2.5-0.5B, to encode coarse scene semantics, including interaction patterns, scene intent, and high-level motion cues. Its hidden states are projected into the semantic feature space to obtain  $F_{\text{sem}}$ , avoiding the overhead of explicit text decoding and enabling efficient contextual guidance for downstream decision making. To strengthen the semantic capability of the compact model, we further employ a CoT-based latent distillation paradigm during training. Specifically, the teacher model receives the same scene information  $\mathcal{I}$ , serialized as a structured textual scene description, together with a handcrafted CoT prompt that guides causal analysis of topology, interaction risk, agent intent, and high-level strategy. The teacher produces both an explicit CoT response for semantic supervision and final-layer hidden representations for latent alignment. The student model is trained on the same structured scene input, and its final-layer hidden states are mean-pooled and projected into the teacher feature space for alignment. In this way, the compact LM learns not only to fit the teacher-augmented semantic targets, but also to approximate the teacher’s internal reasoning geometry in latent space. As a result, the fast-thinking pathway can encode richer causal and contextual information while preserving low inference cost. Full details regarding this distillation paradigm are provided in Section 3.5.

### 3.4.3. Dual-Pathway Cross-modal Attention

To effectively fuse semantic and geometric features, we design a dual-pathway cross-modal attention mechanism. As shown in Fig. 2, one pathway injects semantic cues into geometric tokens, while the other uses semantic context to reweight geometric evidence according to scene intent, interaction risk, and motion context. Specifically, let  $F_{\text{geo}}^l \in \mathbb{R}^{N_g \times D}$  denote the geometric features at the  $l$ -th decoder layer, where  $N_g$  is the number of geometric tokens, and  $F_{\text{sem}} \in \mathbb{R}^{N_s \times D}$  denote the semantic latent features, where  $N_s$  is the number of semantic tokens.

**Semantic Injection Pathway.** The first pathway uses geometric motion features as queries to retrieve the most relevant semantic context. Its purpose is to inject scene semantics directly into the motion representation. Formally,

$$Q_{\text{inj}}^l = F_{\text{geo}}^l W_{\text{inj}}^Q, \quad K_{\text{sem}}^l = F_{\text{sem}} W_{\text{sem}}^K, \quad V_{\text{sem}}^l = F_{\text{sem}} W_{\text{sem}}^V \quad (13)$$

Here, the  $W_{\text{inj}}^Q$ ,  $W_{\text{sem}}^K$ , and  $W_{\text{sem}}^V$  are the learnable matrices for the query, key, and value tokens, respectively. For efficiency, we apply low-rank projections to the semantic keys and values along the sequence dimension, yielding  $\tilde{K}_{\text{sem}}^l = (P_{\text{sem}}^l)^\top K_{\text{sem}}^l$  and  $\tilde{V}_{\text{sem}}^l = (R_{\text{sem}}^l)^\top V_{\text{sem}}^l$ , where  $P_{\text{sem}}^l, R_{\text{sem}}^l \in \mathbb{R}^{N_s \times r}$  are learnable projection matrices and  $r \ll N_s$  is the projection rank. The resulting features are then used to compute the semantic-injection output:

$$H_{\text{inj}}^l = \phi_{\text{Attn}}(Q_{\text{inj}}^l, \tilde{K}_{\text{sem}}^l, \tilde{V}_{\text{sem}}^l) \quad (14)$$

where  $\phi_{\text{Attn}}(Q, K, V) = \phi_{\text{Softmax}}(QK^\top / \sqrt{D})V$  denotes scaled dot-product attention.

**Geometric Grounding Pathway.** Next, the second pathway uses semantic context to query the geometric tokens and identify the most relevant physically grounded motion evidence. Instead of attending to all semantic tokens independently, we first summarize them into a compact semantic anchor  $\mathbf{a}_{\text{sem}}^l = \phi_{\text{Pool}}(F_{\text{sem}}) \in \mathbb{R}^{1 \times D}$ , and then map semantic anchors  $\mathbf{a}_{\text{sem}}^l$  and geometric features  $F_{\text{geo}}^l$  to query, key, and value tokens, respectively:

$$Q_{\text{grd}}^l = \mathbf{a}_{\text{sem}}^l W_{\text{grd}}^Q, \quad K_{\text{geo}}^l = F_{\text{geo}}^l W_{\text{geo}}^K, \quad V_{\text{geo}}^l = F_{\text{geo}}^l W_{\text{geo}}^V \quad (15)$$

where the  $W_{\text{grd}}^Q$ ,  $W_{\text{geo}}^K$ , and  $W_{\text{geo}}^V$  are also the learnable matrices. Similarly, we apply low-rank projections to the geometric keys and values, yielding  $\tilde{K}_{\text{geo}}^l = (P_{\text{geo}}^l)^\top K_{\text{geo}}^l$  and  $\tilde{V}_{\text{geo}}^l = (R_{\text{geo}}^l)^\top V_{\text{geo}}^l$ , where  $P_{\text{geo}}^l, R_{\text{geo}}^l \in \mathbb{R}^{N_g \times r}$ .

Based on these projected features, the grounding pathway produces a geometry-aware semantic context, denoted by  $\mathbf{c}_{\text{grd}}^l = \phi_{\text{Attn}}(\mathbf{Q}_{\text{grd}}^l, \tilde{\mathbf{K}}_{\text{geo}}^l, \tilde{\mathbf{V}}_{\text{geo}}^l) \in \mathbb{R}^{1 \times D}$ . This context vector is then broadcast to the geometric token dimension as  $\mathbf{C}_{\text{grd}}^l = \mathbf{1}_{N_g} \mathbf{c}_{\text{grd}}^l \in \mathbb{R}^{N_g \times D}$ , so that it can be fused with the token-wise semantic injection features in a unified representation. Following this, the two pathways are combined through a learnable gate. Formally,

$$\Gamma^l = \sigma\left(\phi_{\text{MLP}}([\mathbf{F}_{\text{geo}}^l \parallel \mathbf{H}_{\text{inj}}^l \parallel \mathbf{C}_{\text{grd}}^l])\right) \quad (16)$$

Where the  $\sigma(\cdot)$  is the sigmoid function, and the fused cross-modal representation  $\mathbf{F}_{\text{cro}}^{l+1}$  is updated as follows:

$$\begin{aligned} \mathbf{F}_{\text{cro}}^{l+1} &= \phi_{\text{LN}}\left(\mathbf{F}_{\text{geo}}^l + \Gamma^l \odot \mathbf{H}_{\text{inj}}^l + (1 - \Gamma^l) \odot \mathbf{C}_{\text{grd}}^l\right) \\ &= \phi_{\text{LN}}\left(\mathbf{F}_{\text{geo}}^l + \Gamma^l \odot \phi_{\text{Attn}}\left(\mathbf{Q}_{\text{inj}}^l, \tilde{\mathbf{K}}_{\text{sem}}^l, \tilde{\mathbf{V}}_{\text{sem}}^l\right) + (1 - \Gamma^l) \odot \phi_{\text{Attn}}\left(\mathbf{Q}_{\text{grd}}^l, \tilde{\mathbf{K}}_{\text{geo}}^l, \tilde{\mathbf{V}}_{\text{geo}}^l\right)\right) \end{aligned} \quad (17)$$

where  $\phi_{\text{LN}}(\cdot)$  is layer normalization. The first term preserves geometric structure through residual fusion, the second injects token-level semantic cues, and the third provides scene-level semantic grounding. Compared with a standard full attention mechanism, whose complexity scales as  $O(N_g N_s D)$ , the low-rank formulation reduces the dominant attention cost to  $O((N_g + N_s)rD)$ , where  $r \ll \min(N_g, N_s)$ . By stacking  $L$  such layers, the model progressively refines the cross-modal representation, and the final fused cross-modal features are denoted by  $\mathbf{F}_{\text{cro}} = \mathbf{F}_{\text{cro}}^L$ .

#### 3.4.4. Cross-modal Decoder

Guided by the cross-modal features  $\mathbf{F}_{\text{cro}}$ , the decoder jointly refines the multimodal predictions for surrounding agents and the planning for the ego agent. For each surrounding agent  $i$ , the decoder leverages its cross-modal feature  $\mathbf{f}_{\text{cro},i}$  to update the initial GMM parameters  $\Theta_i$  generated by the STE. Specifically, it predicts a residual correction  $\Delta\Theta_i = \phi_{\text{pred}}(\mathbf{f}_{\text{cro},i})$  to yield the refined parameters  $\bar{\Theta}_i = \Theta_i + \Delta\Theta_i$ . The final predictive distribution is formulated as  $P(\mathbf{Y}_i | \mathbf{F}_{\text{cro}}) = \sum_{k=1}^K \bar{\pi}_{i,k} \mathcal{N}(\mathbf{Y}_i | \bar{\mu}_{i,k}, \bar{\Sigma}_{i,k})$ . This residual update mechanism ensures that high-level semantic reasoning refines, rather than replaces, the original kinematic prior. Consequently, it produces the final future trajectories  $\mathbf{Y}_{1:n}^{t+1:t+T_f}$  grounded in both physical limits and interaction semantics. For planning, we employ an anchor-based refinement strategy on the initial planning proposals  $\mathcal{T}_{\text{ego}} = \{\tau_k^*\}_{k=1}^K$ , avoiding the instability of generating trajectories from scratch. For each candidate  $\tau_k^*$ , the decoder extracts a proposal-aware feature  $\mathbf{c}_k = \phi_{\text{plan}}(\mathbf{F}_{\text{cro}}, \tau_k^*)$ . This feature simultaneously predicts a trajectory correction  $\Delta\tau_k = \phi_{\text{ref}}(\mathbf{c}_k)$  and a confidence score  $s_k = \phi_{\text{score}}(\mathbf{c}_k)$ . The refined proposal is computed as  $\bar{\tau}_k = \tau_k^* + \Delta\tau_k$ . Finally, the optimal planned trajectory  $\mathbf{Y}_0^{t+1:t+T_f}$  is selected as the highest-scoring candidate  $\bar{\tau}_{k^*}$ , where  $k^* = \arg \max_k s_k$ .

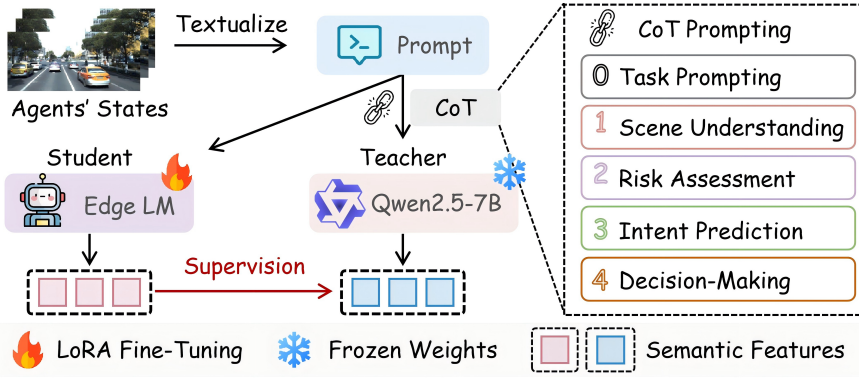
### 3.5. Training and Inference

The training procedure of LLM-TP<sup>2</sup> is divided into two stages. In the first stage, we train the compact language model used in the fast-thinking mode through LoRA fine-tuning and our proposed CoT-based latent distillation, empowering the lightweight edge LM with the structural reasoning and contextual priors of a massive multimodal LLM. In the second stage, we optimize the prediction-and-planning network using trajectory and planning supervision, while initializing the semantic branch with the distilled compact LM obtained from the first stage.

#### 3.5.1. Stage-1: LoRA Fine-Tuning and CoT-based Latent Distillation

The objective of this stage is to adapt the compact LM to the autonomous-driving domain through efficient LoRA fine-tuning and further endow it with stronger causal scene understanding through CoT-based latent distillation from a stronger teacher LLM. Technically, we begin with LoRA fine-tuning on the paired scene-text data from the Highway-text and Urban-text datasets. Let  $\mathcal{Q}$  denote the input state space of a traffic scene, and  $\mathcal{B} = \{b_1, \dots, b_{|\mathcal{B}|}\}$  denote its corresponding textual supervision sequence provided by the datasets. Given  $(\mathcal{I}, \mathcal{B})$ , the student model (edge LM)  $f_{\text{stu}}$  is trained under the standard autoregressive objective to model the scene semantics. This step allows the compact LM to acquire domain-specific parsing ability, such as recognizing map topology, traffic participants, and common interaction patterns. However, standard autoregressive fine-tuning mainly encourages the model to fit the target text distribution, and therefore does not explicitly transfer the causal reasoning process required for complex driving scenarios. To address this, we further introduce a CoT-based latent distillation paradigm that combines LoRA fine-tuning with latent-space alignment. For each scene, a frozen teacher model  $f_{\text{tea}}$  (Qwen2.5-VL-7B) is prompted with a structured CoT





**Figure 3:** Illustration of the proposed CoT-based distillation paradigm. A high-capacity teacher model (Qwen2.5-VL-7B) processes the driving scene to generate a structured causal reasoning sequence, explicitly analyzing environmental topology, potential risks, agent intents, and strategic decisions. Rather than relying solely on explicit textual supervision, the framework aligns the projected hidden states of the lightweight student model directly with the latent feature anchor of the teacher. The internal representations of the lightweight student language model are then mapped through a cognitive projection head to align with this anchor. Combined with LoRA fine-tuning, this dual alignment strategy aims to endow the compact student model with profound semantic reasoning while maintaining optimal inference efficiency.

template to generate a step-by-step causal interpretation of the scene. As shown in Fig. 3, the prompt explicitly guides the teacher model to: (1) analyze the environmental topology and map constraints, (2) identify potential collision risks and their causes, (3) infer the future intent and motion evolution of critical agents, and (4) derive a high-level strategic decision for the ego vehicle. Let  $\tilde{\mathcal{B}} = [\mathcal{C}; \mathcal{A}]$  denote the teacher model's augmented supervision sequence, obtained by concatenating the generated reasoning chain  $\mathcal{C}$  and the final semantic answer  $\mathcal{A}$ . Here,  $\tilde{\mathcal{B}}$  is used as the distilled semantic target for student training, thereby enriching the original dataset annotation with explicit causal reasoning.

Beyond textual supervision, we also extract a latent teacher feature anchor from the final hidden states of the teacher model. Specifically, the teacher feature is defined as  $H_{\text{tea}} = \phi_{\text{Pool}}(f_{\text{tea}}(\mathcal{I}, \mathcal{P}_{\text{cot}})) \in \mathbb{R}^{d_t}$ , where  $\mathcal{P}_{\text{cot}}$  denotes the CoT prompt and  $\phi_{\text{Pool}}(\cdot)$  represents mean-pooling over the final hidden sequence. This latent anchor encodes the teacher's internal reasoning geometry over scene structure, interaction risk, agent intent, and strategic planning cues.

Given the same scene information  $\mathcal{Q}$ , the student model produces a pooled hidden representation, which can be denoted as  $H_{\text{stu}} = \phi_{\text{Pool}}(f_{\text{stu}}(\mathcal{I}; \theta_0, \Delta\theta)) \in \mathbb{R}^{d_s}$  where  $\theta_0$  denotes the frozen pretrained backbone parameters and  $\Delta\theta$  denotes the trainable LoRA parameters. Since  $d_s \neq d_t$ , we introduce a lightweight cognitive projection head  $P_{\omega}(\cdot)$ , implemented by a GRU followed by an MLP, to map the student feature into the teacher feature space, i.e.,  $\hat{H}_{\text{stu}} = P_{\omega}(H_{\text{stu}}) \in \mathbb{R}^{d_t}$ . This projected feature is then aligned with the teacher anchor  $H_{\text{tea}}$  in latent space.

Overall, the loss function in stage-1 consists of two components. The first is the LoRA fine-tuning loss:

$$\mathcal{L}_{\text{LoRA}} = -\frac{1}{|\tilde{\mathcal{B}}|} \sum_{i=1}^{|\tilde{\mathcal{B}}|} \log p_{\text{stu}}(\tilde{b}_i | \tilde{b}_{<i}, \mathcal{I}; \theta_0, \Delta\theta) \quad (18)$$

where  $\tilde{b}_i$  denotes the  $i$ -th token in the teacher-augmented sequence  $\tilde{\mathcal{B}}$ , and  $p_{\text{stu}}(\cdot)$  is the token-level conditional probability distribution induced by the student LM. This term encourages the compact model to fit the teacher-enriched semantic supervision in the vocabulary space. The second is the latent distillation loss, which can be represented as:

$$\mathcal{L}_{\text{CoT}} = \left\| \hat{H}_{\text{stu}} - H_{\text{tea}} \right\|_2^2 \quad (19)$$

which forces the internal representation of the student to approach the hidden reasoning feature of the teacher. The final Stage-1 objective is formally denoted as follows:

$$\mathcal{L}_{\text{stage-1}} = \mathcal{L}_{\text{LoRA}} + \lambda_{\text{cot}} \mathcal{L}_{\text{CoT}} \quad (20)$$

where  $\lambda_{\text{cot}}$  controls the strength of latent reasoning transfer. During optimization, the pretrained backbone remains frozen, while the LoRA parameters  $\Delta\theta$  and the projection-head parameters  $\omega$  are jointly updated. In this way, the

compact model learns not only to reproduce semantically meaningful text, but also to align its hidden feature geometry with the teacher's causal reasoning structure under the same driving context. After training, the optimized hidden representation is further mapped through a lightweight semantic adapter  $\phi_{\text{sem}}(\cdot)$  to obtain the latent semantic feature  $F_{\text{sem}} = \phi_{\text{sem}}(H_{\text{stu}})$ , which is subsequently used by the ARE in the fast-thinking mode. Compared with conventional fine-tuning, which only supervises the final textual output, the proposed paradigm explicitly transfers reasoning structure in the latent space. As a result, the LoRA-adapted compact LM acquires stronger scene understanding and more reliable causal abstraction, while retaining the low inference cost required for AD deployment.

### 3.5.2. Stage-2: Joint Training for Trajectory Prediction and Planning

Building upon the semantic alignment achieved in the stage-1, this stage jointly optimizes the Cross-modal Decoder for multimodal trajectory prediction and safety-aware ego-vehicle planning. The total loss function for this stage consists of two components: (1) Prediction loss and (2) Planning loss. Formally,

$$\mathcal{L}_{\text{stage-2}} = \lambda_{\text{prediction}} \mathcal{L}_{\text{pre}} + \lambda_{\text{plan}} \mathcal{L}_{\text{planning}} \quad (21)$$

where  $\lambda_{\text{prediction}}$  and  $\lambda_{\text{plan}}$  are hyperparameters balancing the contributions of the prediction and the planning tasks.

**Prediction Loss.** To model the multimodal behaviors of surrounding agents, we represent each neighboring trajectory distribution using a  $K$ -component GMM, where the optimal mode is selected through negative log-likelihood (NLL) minimization based on the predicted means, covariances, and trajectory positions. Since different datasets follow different evaluation conventions, we define the trajectory prediction loss in a dataset-aware manner to ensure fair comparison with prior methods. For the HighD and NGSIM datasets, we follow the loss design commonly adopted in prior work [44], combining the mean squared error loss  $\mathcal{L}_{\text{MSE}}$  with the NLL loss  $\mathcal{L}_{\text{NLL}}$ . For the nuScenes dataset, we pair  $\mathcal{L}_{\text{NLL}}$  with the average displacement error loss  $\mathcal{L}_{\text{ADE}}$ . The overall prediction loss is defined as follows:

$$\mathcal{L}_{\text{prediction}} = \begin{cases} \mathcal{L}_{\text{NLL}} + \mathcal{L}_{\text{ADE}}, & \text{for nuScenes} \\ \mathcal{L}_{\text{NLL}} + \mathcal{L}_{\text{MSE}}, & \text{for HighD and NGSIM} \end{cases} \quad (22)$$

**Planning Loss.** For ego-agent planning, we adopt a lightweight supervision strategy that is consistent with the proposal-based decoder. Specifically, the final planned trajectory  $Y_0^{t+1:t+T_f}$  is directly supervised against the ground-truth future path  $Y_0^{t+1:t+T_f}$  using a smooth  $\mathcal{L}_1$  loss. In addition, since the planner predicts a confidence score for each candidate proposal, we further supervise the proposal-ranking branch with a cross-entropy loss  $\mathcal{L}_{\text{cls}}$ . To construct the classification target, we select the proposal whose refined trajectory is closest to the ground-truth path and use its index as the positive label. The planning loss is therefore defined as follows:

$$\mathcal{L}_{\text{planning}} = \mathcal{L}_1(Y_0^{t+1:t+T_f}, Y_0^{t+1:t+T_f}) + \lambda_{\text{cls}} \mathcal{L}_{\text{cls}}(\mathbf{s}, \mathbf{s}^*), \quad (23)$$

where  $\mathbf{s}$  denotes the predicted proposal scores,  $\mathbf{s}^*$  is the one-hot target derived from the oracle proposal, and  $\lambda_{\text{cls}}$  balances trajectory regression and score supervision. This formulation keeps the planning objective simple while still encouraging the decoder to produce both an accurate final path and a reliable proposal ranking.

### 3.5.3. Inference

During inference, the framework supports both scene annotation generation and joint trajectory prediction and planning. For annotation generation, the historical agent states are first serialized into structured text and then fed into the distilled edge LM to generate semantic descriptions under prompt guidance. For the trajectory prediction and planning task, the teacher model does not participate in inference. The Hierarchical Surprise Monitor estimates the scene surprise and produces a routing signal, based on which the Adaptive Reasoning Engine selects either the frozen compact LM or the stronger frozen LLM for semantic reasoning. Importantly, we do not use the generated text directly. Instead, the hidden representations of the selected language model are mapped to the semantic latent feature  $F_{\text{sem}}$ . This feature is then fused with the geometric representation  $F_{\text{geo}}$  from the STE to obtain cross-modal features  $F_{\text{cro}}$ , which are used to refine surrounding-agent prediction and generate the final planned trajectory for the autonomous vehicle.

## 4. Experiments

### 4.1. Experiment Setup

**Datasets.** To comprehensively validate our proposed framework, we design a multi-tiered empirical evaluation protocol. Initially, the lightweight language models are fine-tuned on our custom Highway-Text and Urban-Text corpora to rigorously assess their semantic scene comprehension. Subsequently, to evaluate trajectory forecasting efficacy, we benchmark our model against state-of-the-art approaches using three prominent real-world datasets: NGSIM[45], HighD [46], and nuScenes [47]. This diverse testing paradigm ensures robust performance across highly heterogeneous traffic topologies, ranging from dense, multi-agent urban intersections to complex campus layouts and high-speed expressways. Furthermore, to validate the downstream trajectory planning capabilities, we execute rigorous closed-loop simulations utilizing the standardized nuPlan [10] environment. These datasets are described as follows:

- *NGSIM*: Initiated by the U.S. Department of Transportation, the Next Generation Simulation (NGSIM) dataset provides highly detailed vehicle trajectory data, including position coordinates and speeds. The data was meticulously collected via overhead camera systems at a sampling rate of 10Hz. The data was collected from the eastbound lanes of Interstate 80 (I-80) in the San Francisco Bay Area and the southbound lanes of Interstate 101 (I-101) in Los Angeles, capturing authentic highway driving conditions.
- *HighD*: Recorded using unmanned aerial vehicles (drones) hovering over a 420-meter section of the German Autobahn, the HighD dataset provides high-precision aerial trajectory data without the occlusion issues typical of ground-level sensors. Sampled at 25Hz between 2017 and 2018, this extensive dataset captures the naturalistic driving maneuvers of over 110,000 vehicles, which encompasses both commercial trucks and passenger cars, spanning a cumulative driven distance of approximately 45,000 kilometers.
- *nuScenes*: Recognized as a pioneering large-scale benchmark for AD perception and prediction, nuScenes comprises 1,000 driving logs recorded in the complex urban areas of Boston and Singapore. Each 20-second scene is meticulously annotated with 3D bounding boxes at 2Hz. The dataset is equipped with a comprehensive 360-degree multimodal sensor suite, which features six cameras, five radars, and one 32-beam LiDAR, yielding 28,130 training, 6,019 validation, and 6,008 testing samples for robust spatial-temporal modeling.
- *nuPlan*: Following the success of nuScenes, nuPlan is the world's first large-scale machine learning benchmark specifically designed for autonomous vehicle planning. It significantly scales up the complexity of driving datasets, containing over 1,300 hours of high-quality sensor data collected across diverse topographies and traffic conditions in four cities (Boston, Pittsburgh, Las Vegas, and Singapore). Unlike traditional perception-focused datasets, nuPlan provides rich, high-definition map data, traffic light states, and detailed multi-agent trajectories, explicitly supporting both open-loop motion forecasting and closed-loop reactive planning evaluation.
- *Highway-Text & Urban-Text*: These two datasets contain over 10 million words of traffic-scene descriptions [17]. Highway-Text comprises 6,606 highway scenarios derived from NGSIM and HighD, while Urban-Text contains 9,310 diverse urban scenarios, including intersections, roundabouts, and campus-road environments, collected from MoCAD [44], ApolloScape [48], and nuScenes. The semantic annotations are generated using GPT-4 Turbo through a structured CoT prompting procedure that progressively performs background and traffic-state analysis, interaction analysis, risk assessment, and maneuver prediction. The outputs from these stages are subsequently consolidated into standardized scene descriptions. Nevertheless, as with any LLM-assisted annotation process, residual semantic noise and occasional inconsistencies may remain. Accordingly, these annotations are used as auxiliary semantic supervision rather than being treated as error-free ground-truth labels.

**Data Segmentation.** For the NGSIM and HighD datasets, continuous trajectory recordings are partitioned into 8-second segments. During evaluation on these specific datasets, HD map priors are explicitly excluded from the model input. Within each segment, the initial 3 seconds serve as the historical observation horizon ( $t_h = 3$ ), while the subsequent 5 seconds formulate the ground-truth prediction horizon ( $t_f = 5$ ). For the nuScenes dataset, we align with standard benchmarking protocols to ensure fair comparisons against existing baselines, adopting a 2-second observation history ( $t_h = 2$ ) to forecast a 4-second future trajectory ( $t_f = 4$ ). To assess actionable decision-making, our closed-loop planning evaluations utilize the Val14 subset of the nuPlan benchmark. This subset encompasses 1,090 diverse scenarios that are individually simulated at a 10 Hz frequency over a 15-second duration. These simulations provide a rich contextual scene complete with HD maps, multi-modal sensor data, and multi-agent trajectories.

**Table 1**

Evaluation of our model and baselines on the NGSIM over a 5-second horizon. Performance is measured by RMSE (m). Unavailable baseline values are marked with ('-'). Top performances are **bolded**, while runner-up results are underlined.

| Model               | Prediction Horizon (s) |             |             |             |             |
|---------------------|------------------------|-------------|-------------|-------------|-------------|
|                     | 1                      | 2           | 3           | 4           | 5           |
| S-GAN [50]          | 0.57                   | 1.32        | 2.22        | 3.26        | 4.40        |
| CS-LSTM [51]        | 0.61                   | 1.27        | 2.09        | 3.10        | 4.37        |
| MATF-GAN [52]       | 0.66                   | 1.34        | 2.08        | 2.97        | 4.13        |
| DRBP [53]           | 1.18                   | 2.83        | 4.22        | 5.82        | –           |
| MFP [54]            | 0.54                   | 1.16        | 1.89        | 2.75        | 3.78        |
| NLS-LSTM [55]       | 0.56                   | 1.22        | 2.02        | 3.03        | 4.30        |
| MHA-LSTM [56]       | 0.41                   | 1.01        | 1.74        | 2.67        | 3.83        |
| WSiP [5]            | 0.56                   | 1.23        | 2.05        | 3.08        | 4.34        |
| CF-LSTM [57]        | 0.55                   | 1.10        | 1.78        | 2.73        | 3.82        |
| STDAN [58]          | 0.42                   | 1.01        | 1.69        | 2.56        | 3.67        |
| iNATran [59]        | 0.39                   | 0.96        | 1.61        | 2.42        | 3.43        |
| FHIF [60]           | 0.40                   | 0.98        | 1.66        | 2.52        | 3.63        |
| GaVa [61]           | 0.40                   | 0.94        | 1.52        | 2.24        | 3.13        |
| DACR-AMTP [62]      | 0.57                   | 1.07        | 1.68        | 2.53        | 3.40        |
| HiT [63]            | <u>0.38</u>            | 0.90        | <u>1.42</u> | 2.08        | 2.87        |
| DEMO [64]           | <b>0.36</b>            | <u>0.86</u> | 1.48        | 2.10        | 2.88        |
| LLM-TP <sup>2</sup> | <u>0.38</u>            | <b>0.83</b> | <b>1.37</b> | <b>1.85</b> | <b>2.63</b> |

**Evaluation Metrics.** Our evaluation protocol spans three hierarchical tasks: (1) semantic reasoning, (2) open-loop trajectory forecasting, and (3) closed-loop planning. First, the semantic scene comprehension of the LM is assessed via BERT score [49]. By computing Precision, Recall, and F1 scores, we quantitatively assess the linguistic alignment between the generated outputs and the reference annotations. Second, for open-loop forecasting, we utilize Root Mean Square Error (RMSE) to evaluate spatial accuracy on the NGSIM and HighD datasets. Conversely, on the nuScenes dataset, we report the minimum Average Displacement Error ( $\min ADE_k$ ,  $k \in \{1, 5, 10\}$ ), Final Displacement Error (FDE), and Miss Rate ( $MR_k$ ,  $k \in \{5, 10\}$ ) to capture predictive performance. Finally, downstream trajectory planning is validated using the official nuPlan evaluation metrics. We emphasize the Closed-Loop Simulation Non-Reactive (CLS-NR) score, a comprehensive metric that evaluates real-world driving proficiency by aggregating multiple penalties. These include target at-fault Collisions (excluding those initiated by external agents, S-CR), Time-to-Collision (S-TTC) boundary violations, drivable area transgressions (S-Area), and deviations from mandated traffic directions. Furthermore, the CLS-NR incorporates continuous evaluations of kinematic comfort (bounding accelerations, jerks, and yaw rates), speed limit compliance, and overall route progress (S-PR) relative to an expert human driver.

**Implementation Details.** The proposed LLM-TP<sup>2</sup> framework is implemented using PyTorch and trained on a cluster of four NVIDIA A40 (48 GB) GPUs. To transfer the reasoning capabilities of the Teacher LLM to the lightweight candidate models without inducing catastrophic forgetting, we employ supervised fine-tuning augmented with LoRA. During this phase, the models are trained using bf16 mixed precision to ensure computational efficiency. The optimization is performed over 10 epochs with a batch size of 8. We utilize an initial learning rate of  $2 \times 10^{-5}$ , complemented by a linear warm-up over the first 500 steps and a weight decay factor of 0.01. This configuration allows the student LMs to effectively clone the step-by-step semantic rationales without updating the full parameter space.

## 4.2. Open-loop Trajectory Prediction Results

**Navigating High-Dynamic Highway Scenes.** Highway scenarios present unique predictive challenges characterized by elevated vehicular velocities and rapid, decisive lane-changing maneuvers. In these environments, even minor miscalculations in short-term intent prediction can compound into severe long-term spatial displacement errors. We evaluate our proposed LLM-TP<sup>2</sup> framework against state-of-the-art baselines on two prominent highway benchmarks: NGSIM and HighD. Specifically, Table 1 presents the predictive performance on the NGSIM dataset, a benchmark notorious for its dense, interactive traffic flows. As the prediction horizon extends from 1 to 5 seconds, traditional kinematic and attention-based models (WSiP, STDAN) exhibit a rapid escalation in RMSE. Compared to the most competitive SOTA baselines (HiT and DEMO), our framework achieves an 11.1% reduction in Root Mean Square Error

**Table 2**

Evaluation of our model and baselines on the HighD over a 5-second horizon. Performance is measured by RMSE (m). Unavailable baseline values are marked with ('-'). Top performances are **bolded**, while runner-up results are underlined.

| Model                     | Prediction Horizon (s) |             |             |             |             |
|---------------------------|------------------------|-------------|-------------|-------------|-------------|
|                           | 1                      | 2           | 3           | 4           | 5           |
| WSiP [5]                  | 0.20                   | 0.60        | 1.21        | 2.07        | 3.14        |
| MHA-LSTM [56]             | 0.19                   | 0.55        | 1.10        | 1.84        | 2.78        |
| NLS-LSTM [55]             | 0.20                   | 0.57        | 1.14        | 1.90        | 2.91        |
| DRBP [53]                 | 0.41                   | 0.79        | 1.11        | 1.40        | –           |
| EA-Net [65]               | 0.15                   | 0.26        | 0.43        | 0.78        | 1.32        |
| CF-LSTM [57]              | 0.18                   | 0.42        | 1.07        | 1.72        | 2.44        |
| STDAN [58]                | 0.19                   | 0.27        | 0.48        | 0.91        | 1.66        |
| DACR-AMTP [62]            | 0.10                   | 0.17        | 0.31        | 0.54        | 1.01        |
| GaVa [61]                 | 0.17                   | 0.24        | 0.42        | 0.86        | 1.31        |
| HiT [63]                  | 0.08                   | <u>0.13</u> | 0.22        | <u>0.39</u> | <u>0.61</u> |
| DEMO [64]                 | <u>0.06</u>            | 0.14        | 0.25        | 0.44        | 0.70        |
| <b>LLM-TP<sup>2</sup></b> | <b>0.05</b>            | <b>0.11</b> | <b>0.15</b> | <b>0.23</b> | <b>0.48</b> |

**Table 3**

Performance comparison of our model and other baselines on the nuScenes dataset. Cases marked as ('-') indicate unspecified values. **Bold** and underlined values represent the best and second-best performance, respectively.

| Model                         | Venue         | minADE <sub>10</sub> ↓ | minADE <sub>5</sub> ↓ | minADE <sub>1</sub> ↓ | FDE ↓       |
|-------------------------------|---------------|------------------------|-----------------------|-----------------------|-------------|
| MTP [66]                      | ICRA'20       | 1.74                   | 2.22                  | 4.42                  | 10.36       |
| MultiPath [67]                | CoRL'20       | 1.14                   | 1.44                  | 3.16                  | 7.69        |
| CoverNet [68]                 | CVPR'20       | 1.92                   | 2.62                  | –                     | 11.36       |
| Trajectron++ [69]             | ECCV'20       | 1.51                   | 1.88                  | –                     | 9.52        |
| AgentFormer [70]              | ICCV'21       | 1.58                   | 1.97                  | –                     | –           |
| LaPred [71]                   | CVPR'21       | 1.12                   | 1.53                  | 3.51                  | 8.12        |
| STGM [72]                     | IEEE-TITS'22  | 2.34                   | –                     | 3.21                  | 9.62        |
| GoHome [7]                    | ICRA'22       | 1.15                   | 1.42                  | –                     | <u>6.99</u> |
| ContextVAE [73]               | IEEE-RAL'23   | –                      | 1.59                  | 3.54                  | 8.24        |
| EMSIN [6]                     | IEEE-TFS'24   | –                      | 1.77                  | 3.56                  | –           |
| AFormer-FLN [74]              | CVPR'24       | 1.32                   | 1.83                  | –                     | –           |
| E-V <sup>2</sup> -Net-SC [75] | CVPR'24       | 1.13                   | 1.44                  | –                     | –           |
| SeFlow [76]                   | ECCV'24       | <u>0.98</u>            | 1.38                  | –                     | 7.89        |
| AMD [77]                      | ICCV'25       | 1.43                   | <u>1.23</u>           | –                     | <u>6.99</u> |
| WAKE [78]                     | IEEE-TPAMI'25 | 1.10                   | 1.24                  | <u>3.03</u>           | 7.02        |
| <b>LLM-TP<sup>2</sup></b>     | –             | <b>0.95</b>            | <b>1.17</b>           | <b>2.95</b>           | <b>6.85</b> |

(RMSE) at the 4-second horizon from 2.08m to 1.85m, and an 8.4% error reduction at the critical 5-second horizon. These substantial percentage improvements underscore the superior predictive capabilities of LLM-TP<sup>2</sup>, demonstrating that semantic reasoning effectively mitigates the severe error compounding typically observed in complex highway dynamics. To further validate the framework's robustness, we evaluate it on the HighD dataset, as shown in Table 2. Notably, it provides an environment with significantly lower sensor noise due to its drone-captured trajectories, naturally resulting in lower overall RMSE for all models. Interestingly, rather than narrowing the performance gap, this high-fidelity setting actually magnifies the strengths of our approach. LLM-TP<sup>2</sup> maintains a commanding lead across all prediction horizons, yielding an average error reduction of approximately 25.2% relative to the most competitive baselines. This advantage becomes particularly evident in the extended prediction windows. For instance, at the 4-second mark, our framework drives the RMSE down to just 0.23m, representing a striking 41.0% improvement over the strongest baseline, HiT. Similarly, at the 5-second horizon, LLM-TP<sup>2</sup> successfully curtails the spatial deviation



**Table 4**

Evaluation results on nuScenes. **Bold** and underlined values represent the best and second-best results, respectively.

| Method      | Model                         | Publication | minADE <sub>5</sub> ↓ | minADE <sub>10</sub> ↓ | MR <sub>5</sub> ↓ | MR <sub>10</sub> ↓ |
|-------------|-------------------------------|-------------|-----------------------|------------------------|-------------------|--------------------|
| Without LLM | AgentFormer [70]              | ICCV        | 1.86                  | 1.45                   | –                 | –                  |
|             | AFormer-FLN [74]              | CVPR        | 1.83                  | 1.32                   | –                 | –                  |
|             | STGM [72]                     | IEEE-TITS   | –                     | 1.62                   | –                 | –                  |
|             | EMSIN [6]                     | IEEE-TFS    | 1.77                  | –                      | 0.54              | –                  |
|             | E-V <sup>2</sup> -Net-SC [75] | CVPR        | 1.44                  | 1.13                   | –                 | –                  |
|             | GOHOME [7]                    | ICRA        | 1.42                  | 1.15                   | 0.57              | 0.47               |
|             | SeFlow [76]                   | ECCV        | 1.38                  | <u>0.98</u>            | 0.60              | 0.38               |
|             | AutoBots [79]                 | ICLR        | 1.37                  | 1.03                   | 0.62              | 0.44               |
| LLM-based   | VisionTrap [34]               | ECCV        | –                     | 1.48                   | –                 | 0.56               |
|             | Traj-LLM [13]                 | IEEE-TIV    | 1.24                  | 0.99                   | <u>0.41</u>       | <b>0.23</b>        |
| LLM + LM    | <b>LLM-TP<sup>2</sup></b>     | –           | <b>1.17</b>           | <b>0.95</b>            | <b>0.40</b>       | <u>0.31</u>        |

to 0.48m, achieving a 21.3% reduction compared to the next best model’s 0.61m. This highlights a crucial insight: given clean input data, the LLM and LM’s semantic reasoning acts as a powerful regularizer. It restricts long-term predictions to logical driving behaviors, thereby achieving unprecedented forecasting accuracy.

**Navigating Complex Urban Topologies.** The nuScenes benchmark features dense urban environments requiring complex multi-agent navigation. While HD maps provide essential geometric priors, traditional data-driven models often fail to extract the underlying semantic rules and right-of-way negotiations. To address this, LLM-TP<sup>2</sup> fuses precise map geometry with the contextual reasoning of a distilled language model to ensure highly accurate and rule-compliant trajectory forecasts. As shown in Table 3, our framework achieves impressive results against recent map-based baselines such as AMD, WAKE, and SeFlow, consistently minimizing displacement errors. Specifically, it reduces the minADE<sub>5</sub> to 1.17, yielding a 4.9% improvement over the recent AMD model. Moreover, LLM-TP<sup>2</sup> reaches 2.95 of the minADE<sub>1</sub> metric, outperforming WAKE by 2.6%. In long-term end-point prediction, it achieves an FDE of 6.85, surpassing both GoHome and AMD by 2.0%. Furthermore, Table 4 highlights the architectural advantage of our method over pure LLM-based alternatives. While recent text-driven models like Traj-LLM capture general driving logic, they often lack the precise spatial anchoring required for stringent urban navigation. LLM-TP<sup>2</sup> bridges this gap, noticeably outperforming Traj-LLM in key metrics such as minADE<sub>5</sub> (1.17 vs. 1.24) and demonstrating superior predictive stability. These combined results offer a clear insight: pairing geometric maps with linguistic reasoning effectively resolves the fundamental limitations of both traditional spatial models and pure LLMs. Where conventional data-driven approaches struggle to capture complex interactions, LLM-TP<sup>2</sup> applies its CoT-distilled logic to actively interpret road constraints. Grounding semantic reasoning directly in the physical structure of the HD map enables the lightweight LM to handle multi-agent behaviors, such as intersection yielding or lane merging, with exceptional reliability. Ultimately, this fusion significantly enhances prediction robustness in dense urban environments.

### 4.3. Closed-loop Trajectory Planning Results

To comprehensively assess the actionable decision-making capabilities of our proposed model, we conduct rigorous closed-loop simulations using the nuPlan Val14 benchmark. Unlike open-loop forecasting, closed-loop evaluation requires the ego-vehicle to continuously adapt to dynamically unrolling traffic scenarios, serving as a highly realistic proxy for real-world deployment. Specifically, we evaluate the model’s performance in an integrated setup where predicted trajectories directly inform the downstream planning stage to simulate authentic autonomous driving conditions. The evaluation results, categorized by planning paradigms, are detailed in Table 5.

A critical insight derived from these comparisons is the inherent compromise between safety and kinematic smoothness prevalent in conventional architectures. Notably, the pure deep learning methods, such as UrbanDriver and RasterModel, frequently over-optimize for trajectory smoothness at the expense of collision avoidance. For instance, while UrbanDriver achieves a perfect Comfort score of 100.00, it exhibits notable vulnerabilities reflected by a low Collisions score of 85.60. Conversely, rule-based systems like the IDM enforce strict safety thresholds but produce overly conservative maneuvers, limiting their overall Score to 79.31. Hybrid approaches, such as Gameformer, attempt to resolve this dilemma by integrating heuristic logic with neural networks. However, they still lack deep contextual understanding, which caps their performance at an overall score of 82.95. Overcoming these structural bottlenecks,



**Table 5**

Comparative closed-loop planning results on the nuPlan Val14 benchmark. All metrics follow a higher-is-better paradigm. Top performances are highlighted in **bold**, runner-up results are underlined, and third-best results are indicated in **purple**.

| Type            | Planner                    | Score↑ | Collisions↑ | S-TTC↑ | Drivable↑ | Comfort↑ | Progress↑ | Speed↑ |
|-----------------|----------------------------|--------|-------------|--------|-----------|----------|-----------|--------|
| Expert          | Log-Replay                 | 93.68  | 98.76       | 94.40  | 98.07     | 99.27    | 98.99     | 96.47  |
| Rule-based      | IDM [3]                    | 79.31  | 90.92       | 83.49  | 94.04     | 94.40    | 86.16     | 97.33  |
| Pure Learning   | PDM-Open [4]               | 50.24  | 74.54       | 69.08  | 87.89     | 99.54    | 69.86     | 97.72  |
|                 | RasterModel [10]           | 66.92  | 86.97       | 81.46  | 85.04     | 81.46    | 80.60     | 98.03  |
|                 | UrbanDriver [80]           | 67.72  | 85.60       | 80.28  | 90.83     | 100.00   | 80.83     | 91.58  |
|                 | PlanTF [8]                 | 85.30  | 94.13       | 90.73  | 96.79     | 93.67    | 89.83     | 97.78  |
| Hybrid          | Gameformer [30]            | 82.95  | 94.32       | 86.77  | 94.87     | 93.39    | 89.04     | 98.67  |
| Ablation Models | LLM-TP <sup>2</sup> (-ARE) | 84.50  | 94.21       | 88.64  | 95.67     | 93.48    | 89.52     | 98.12  |
|                 | LLM-TP <sup>2</sup> (-HCE) | 77.76  | 89.45       | 82.18  | 93.12     | 91.65    | 84.82     | 95.56  |
|                 | LLM-TP <sup>2</sup> (-CMD) | 85.24  | 94.86       | 89.20  | 96.25     | 94.65    | 89.84     | 98.25  |
| LLM + LM        | LLM-TP <sup>2</sup>        | 90.23  | 97.12       | 92.31  | 97.95     | 99.75    | 91.45     | 98.85  |

our proposed LLM-TP<sup>2</sup> framework introduces a paradigm shift by embedding explicit semantic reasoning directly into the planning pipeline. By capturing the underlying logic of traffic environments, LLM-TP<sup>2</sup> effectively bridges the gap between rigid safety constraints and dynamic adaptability. As shown in the evaluation, our model achieves an exceptional overall Score, outperforming the most competitive pure learning baseline, PlanTF, by approximately 5.8%. Particularly noteworthy is the model's dominance in crucial safety metrics. By surpassing the leading baseline by nearly 3.2% in collision avoidance and 1.7% in S-TTC margins, the distilled LM demonstrates a successful internalization of complex right-of-way rules and interactive yielding behaviors without relying on rigid heuristic coding. Furthermore, our framework does not sacrifice ride quality or efficiency to achieve this safety. Moreover, it maintains a robust Progress score of 91.45 and a stellar Comfort rating of 99.75, delivering a balanced and human-like driving profile. Ultimately, LLM-TP<sup>2</sup> narrows the performance gap with the human expert log-replay (93.68) more effectively than other baselines. These findings confirm that coupling large language models with lightweight planners creates an unprecedented synthesis of safety, compliance, and dynamic efficiency in autonomous driving navigation.

## 4.4. Ablation Studies

### 4.4.1. Ablation Study of Each Component for Trajectory Prediction

To rigorously quantify the contribution of each architectural innovation, we conduct a comprehensive ablation study across the NGSIM, HighD, and nuScenes datasets. Table 6 delineates five ablated variants (Methods A through E) evaluated against our complete LLM-TP<sup>2</sup> framework (Method F). By systematically disabling core modules, we move beyond simply reporting performance drops to uncover the underlying cognitive and physical mechanisms that drive our predictive superiority. Specifically, the most profound degradation occurs when the Hierarchical Contextual Encoder is completely removed (Method C), stripping away all semantic guidance from the language models. Interestingly, while the absolute error penalty is relatively constrained at the 1-second horizon, it compounds drastically as the prediction window extends. At the 5-second mark on the NGSIM dataset, removing the language models inflates the spatial error by a striking 37.6% (from 2.63m to 3.62m). Similarly, Method C suffers a 21.4% deterioration in the minADE<sub>5</sub> metric on nuScenes. This temporal divergence highlights a fundamental principle of autonomous navigation. Short-term trajectories are heavily governed by physical inertia, making purely kinematic models temporarily viable. Conversely, long-term forecasting demands a deep understanding of human intent, right-of-way negotiation, and complex multi-agent interactions, proving that explicit semantic reasoning acts as an indispensable cognitive anchor.

Moreover, Method D evaluates the impact of disabling the Adaptive Reasoning Engine by restricting the model to an LM that relies solely on the fast thinking mode. Without the deep reasoning capacity provided by the high-capacity LLM, the model lacks the semantic depth required to resolve complex causal interactions and safety-critical edge cases. This monolithic approach degrades the minADE<sub>5</sub> accuracy on nuScenes by approximately 14.5%, confirming the necessity of our dual process architecture. These results highlight that while the fast thinking mode ensures real-time efficiency for routine maneuvers, the selective invocation of deep thinking reasoning is essential for maintaining predictive precision in highly uncertain environments. Method E replaces the cross-modal attention mechanism with a standard self-attention mechanism, effectively severing the explicit alignment between high-level semantic features

**Table 6**

Architectural configurations of the evaluated variants in our ablation study. Methods A through E represent ablated versions with specific modules removed, while Method F denotes the complete proposed framework. The symbols ✓ and ✗ indicate the inclusion and exclusion of the respective components.

| Components                      | Ablation Methods |   |   |   |   |   |
|---------------------------------|------------------|---|---|---|---|---|
|                                 | A                | B | C | D | E | F |
| Spatial Interaction Modeling    | ✗                | ✓ | ✓ | ✓ | ✓ | ✓ |
| Temporal Dynamics Modeling      | ✓                | ✗ | ✓ | ✓ | ✓ | ✓ |
| Hierarchical Contextual Encoder | ✓                | ✓ | ✗ | ✓ | ✓ | ✓ |
| Adaptive Reasoning Engine       | ✓                | ✓ | ✓ | ✗ | ✓ | ✓ |
| Cross-modal Decoder             | ✓                | ✓ | ✓ | ✓ | ✗ | ✓ |

**Table 7**

Trajectory-prediction errors (RMSE in m) of different ablated variants on the NGSIM and HighD datasets. The lowest error at each prediction horizon is **bolded** and corresponds to the complete LLM-TP<sup>2</sup> framework.

| Dataset | Time (s) | Ablation Methods |      |      |      |      |             |
|---------|----------|------------------|------|------|------|------|-------------|
|         |          | A                | B    | C    | D    | E    | F           |
| NGSIM   | 1        | 0.41             | 0.40 | 0.54 | 0.39 | 0.44 | <b>0.38</b> |
|         | 2        | 0.98             | 0.95 | 1.21 | 0.96 | 1.08 | <b>0.83</b> |
|         | 3        | 1.45             | 1.50 | 1.78 | 1.54 | 1.61 | <b>1.37</b> |
|         | 4        | 2.48             | 2.53 | 2.99 | 2.17 | 2.58 | <b>1.85</b> |
|         | 5        | 3.10             | 3.17 | 3.62 | 3.16 | 3.25 | <b>2.63</b> |
| HighD   | 1        | 0.07             | 0.07 | 0.08 | 0.06 | 0.06 | <b>0.05</b> |
|         | 2        | 0.15             | 0.14 | 0.21 | 0.13 | 0.15 | <b>0.11</b> |
|         | 3        | 0.20             | 0.21 | 0.35 | 0.18 | 0.25 | <b>0.15</b> |
|         | 4        | 0.30             | 0.29 | 0.52 | 0.33 | 0.42 | <b>0.23</b> |
|         | 5        | 0.57             | 0.54 | 0.81 | 0.61 | 0.69 | <b>0.48</b> |

**Table 8**

Ablation results on the nuScenes dataset. **Bold** values represent the results of the complete proposed framework.

| Metric               | Ablation Methods |      |      |      |      |             |
|----------------------|------------------|------|------|------|------|-------------|
|                      | A                | B    | C    | D    | E    | F           |
| minADE <sub>5</sub>  | 1.26             | 1.28 | 1.42 | 1.34 | 1.37 | <b>1.17</b> |
| minADE <sub>10</sub> | 1.02             | 1.05 | 1.32 | 1.18 | 1.15 | <b>0.95</b> |
| MR <sub>5</sub>      | 0.47             | 0.44 | 0.65 | 0.53 | 0.50 | <b>0.40</b> |
| MR <sub>10</sub>     | 0.33             | 0.36 | 0.57 | 0.41 | 0.38 | <b>0.31</b> |

and continuous kinematic trajectories. This disconnection leads to a significant 43.7% increase in endpoint deviation on the HighD dataset at the 5-second horizon. Such a sharp performance drop demonstrates that linguistic reasoning alone lacks physical grounding in geometric space. These results confirm that cross-modal fusion is indispensable for translating abstract intents into precise and physically feasible geometric coordinates.

Finally, Methods A and B demonstrate that spatial interaction modeling and temporal dynamics modeling remain fundamental prerequisites for the overall architecture. Disabling the spatial module (Method A) inflates the long-term prediction error on NGSIM by nearly 17.9%, as the model fails to capture concurrent multi-agent geometric constraints. Similarly, removing the temporal modeling component (Method B) yields a 20.5% error increase in the same setting, severely disrupting the continuous physical state tracking of the ego-vehicle and its surroundings. These results confirm a logical baseline expectation. Even when equipped with SOTA linguistic reasoning, the system still requires highly robust kinematic priors to accurately project spatial motion. Ultimately, the synergy of explicit physical modeling, dynamic cognitive routing, and precise cross-modal alignment defines the success of the LLM-TP<sup>2</sup> framework.

**Table 9**

Quantitative assessment of diverse LM backbones on the proposed Highway-Text and Urban-Text datasets. The evaluation reports precision, recall, and F1 score, together with inference latency measured using a batch size of 16. Absolute peak performances are underlined, whereas **bold** values indicate the optimal balance between linguistic accuracy and computational efficiency. The **purple-shaded** row indicates the final configuration selected for our proposed model.

| Model                      | #Param. (B) | CoT-Distill. | Time (s) ↓  | Highway-Text ↑ |               |               | Urban-Text ↑  |               |               |
|----------------------------|-------------|--------------|-------------|----------------|---------------|---------------|---------------|---------------|---------------|
|                            |             |              |             | Precision      | Recall        | F1 Score      | Precision     | Recall        | F1 Score      |
| Qwen2.5-0.5B [81]          | 0.50        | <b>X</b>     | 0.14        | 0.9730         | 0.9735        | 0.9732        | 0.9415        | 0.9420        | 0.9417        |
| Qwen-1.5 [11]              | 0.62        | <b>X</b>     | 0.17        | 0.9736         | 0.9745        | 0.9739        | 0.9427        | 0.9426        | 0.9425        |
| Flan-T5 [82]               | 0.78        | <b>X</b>     | 0.21        | 0.9738         | 0.9748        | 0.9742        | 0.9430        | 0.9435        | 0.9432        |
| TinyLlama [83]             | 1.10        | <b>X</b>     | 0.29        | 0.9739         | 0.9751        | 0.9745        | 0.9408        | 0.9541        | 0.9474        |
| Phi-1.5 [84]               | 1.50        | <b>X</b>     | 0.37        | 0.9767         | 0.9766        | 0.9766        | 0.9433        | 0.9543        | 0.9483        |
| Qwen2.5-1.5B [81]          | 1.50        | <b>X</b>     | 0.38        | 0.9772         | 0.9775        | 0.9773        | 0.9455        | 0.9560        | 0.9505        |
| Qwen-1.5 [11]              | 0.62        | ✓            | 0.21        | 0.9810         | 0.9825        | 0.9817        | 0.9585        | 0.9600        | 0.9592        |
| Flan-T5 [82]               | 0.78        | ✓            | 0.28        | 0.9815         | 0.9830        | 0.9822        | 0.9590        | 0.9610        | 0.9600        |
| TinyLlama [83]             | 1.10        | ✓            | 0.38        | 0.9840         | 0.9855        | 0.9847        | 0.9625        | 0.9645        | 0.9635        |
| Phi-1.5 [84]               | 1.50        | ✓            | 0.48        | 0.9865         | 0.9870        | 0.9867        | 0.9680        | 0.9695        | 0.9687        |
| Qwen2.5-1.5B [81]          | 1.50        | ✓            | 0.49        | 0.9892         | 0.9905        | 0.9898        | 0.9745        | 0.9760        | 0.9752        |
| <b>Qwen2.5-0.5B (Ours)</b> | <b>0.50</b> | ✓            | <b>0.16</b> | <b>0.9845</b>  | <b>0.9852</b> | <b>0.9848</b> | <b>0.9658</b> | <b>0.9670</b> | <b>0.9664</b> |

#### 4.4.2. Ablation Study of Each Component for Trajectory Planning

To further validate the impact of our cognitive architecture on actionable decision-making, we conduct an ablation study within the closed-loop nuPlan benchmark. As shown in Table 5, we systematically evaluate three ablated variants against the full LLM-TP<sup>2</sup> framework to isolate the contributions of semantic understanding, dynamic routing, and cross-modal grounding in continuous driving scenarios. Specifically, evaluating the model without the Hierarchical Contextual Encoder (-HCE) reveals the severe limitations of purely kinematic navigation. Disconnecting the language models results in a 13.8% decrease in the overall driving score. Furthermore, this omission significantly undermines the proactive safety profile of the ego-vehicle. Without semantic priors, the S-TTC metric declines by 11.0%, and the Collisions score falls to 89.45. This sheds light on closed-loop navigation: without explicit semantic reasoning to interpret complex right-of-way rules and anticipate multi-agent interactions, the planner becomes a purely reactive kinematic agent. It fails to foresee hazards, leading to dangerous near-miss scenarios and delayed braking responses.

Furthermore, the efficacy of semantic reasoning heavily depends on its successful translation into precise continuous control commands. Evaluating the architecture without the cross-modal attention mechanism (-CMD) reveals the exact consequences of severing the alignment between linguistic inference and physical trajectory spaces. This ungrounded configuration directly causes a 5.5% contraction in the overall driving Score. Although the framework successfully formulates high-level semantic driving strategies, the absence of precise geometric anchoring inevitably introduces spatial inaccuracies and execution latency. This demonstrates that cognitive intent alone is insufficient for safe closed-loop driving unless it is rigorously fused with physical kinematic states. Ultimately, the removal of the Adaptive Reasoning Engine (-ARE), which relies solely on the language model for reasoning, highlights the importance of dynamic resource allocation in continuous driving tasks. Without the ability to route scenarios based on free energy and scene surprise, the framework is forced to apply a monolithic inference strategy, dropping the overall Score to 84.50. Interestingly, the most notable casualty in this variant is ride quality, with the Comfort metric dropping from a stellar 99.75 to 93.48. This highly specific degradation suggests that a rigid, non-adaptive reasoning approach struggles with execution consistency. It tends to overreact in trivial traffic flows or under-process highly interactive edge cases, resulting in jerky, unrefined, and uncomfortable vehicle maneuvers. Overall, these results confirm that human-level closed-loop navigation cannot be achieved through isolated architectural improvements. It demands an integrated framework where high-level semantic reasoning is continuously grounded in precise geometry and dynamically allocated according to scene complexity.

#### 4.4.3. Ablation Study of LM Backbones and CoT-based Distillation

We compare LM backbones with LoRA fine-tuning alone and with additional CoT-based distillation through teacher-guided latent feature alignment. Tables 9 and 10 report the results for text-based scene understanding and downstream trajectory prediction, respectively. In the framework's routing design, the threshold for invoking the

**Table 10**

Ablation comparison of diverse LM backbones within the Hierarchical Contextual Encoder on the nuScenes benchmark. The evaluation reports parameter counts in billions, the effect of CoT distillation, prediction errors, and inference latency. Underlined values indicate the absolute best performance, whereas **bold** values indicate the optimal balance between predictive accuracy and efficiency. The **purple-shaded** row highlights the final configuration adopted for LLM-TP<sup>2</sup>.

| Language Model             | #Param. (B) | CoT-Distill. | Evaluation Metrics ↓ |                      |                 |                  | Latency (ms) ↓ |
|----------------------------|-------------|--------------|----------------------|----------------------|-----------------|------------------|----------------|
|                            |             |              | minADE <sub>5</sub>  | minADE <sub>10</sub> | MR <sub>5</sub> | MR <sub>10</sub> |                |
| Qwen2.5-0.5B (LoRA) [81]   | 0.50        | ✗            | 1.34                 | 1.18                 | 0.53            | 0.41             | 213            |
| Qwen-1.5 (LoRA) [11]       | 0.62        | ✗            | 1.33                 | 1.17                 | 0.52            | 0.40             | 244            |
| Flan-T5 (LoRA) [82]        | 0.78        | ✗            | 1.30                 | 1.12                 | 0.49            | 0.37             | 282            |
| TinyLlama (LoRA) [83]      | 1.10        | ✗            | 1.28                 | 1.09                 | 0.47            | 0.35             | 379            |
| Phi-1.5 (LoRA) [84]        | 1.50        | ✗            | 1.27                 | 1.08                 | 0.46            | 0.34             | 454            |
| Qwen2.5-1.5B (LoRA) [81]   | 1.50        | ✗            | 1.25                 | 1.06                 | 0.44            | 0.33             | 463            |
| Qwen-1.5 (LoRA) [11]       | 0.62        | ✓            | 1.19                 | 0.98                 | 0.42            | 0.33             | 280            |
| Flan-T5 (LoRA) [82]        | 0.78        | ✓            | 1.20                 | 0.99                 | 0.43            | 0.34             | 369            |
| TinyLlama (LoRA) [83]      | 1.10        | ✓            | 1.15                 | 0.93                 | 0.38            | 0.29             | 476            |
| Phi-1.5 (LoRA) [84]        | 1.50        | ✓            | 1.14                 | 0.93                 | 0.39            | 0.28             | 573            |
| Qwen2.5-1.5B (LoRA) [81]   | 1.50        | ✓            | <u>1.11</u>          | <u>0.92</u>          | <u>0.37</u>     | <u>0.26</u>      | 587            |
| <b>Qwen2.5-0.5B (Ours)</b> | <b>0.50</b> | <b>✓</b>     | <b>1.17</b>          | <b>0.95</b>          | <b>0.40</b>     | <b>0.31</b>      | <b>237</b>     |

online Qwen2.5-VL-3B branch is calibrated at the 90th-95th percentiles of validation-set surprise scores, targeting approximately 5-10% of validation scenarios for deeper reasoning. Both latency evaluations use NVIDIA A40 GPUs with an inference batch size of 16. For the text evaluation, each batch contains equal numbers of highway and urban scenarios; for nuScenes, timing covers trajectory prediction and excludes downstream trajectory planning.

On Highway-Text and Urban-Text, CoT-based distillation improves BERTScore F1 across the evaluated backbones. Specifically, Qwen2.5-0.5B achieves relative F1 improvements of 1.2% and 2.6% on the two datasets over its LoRA-only counterpart. Its distilled variant also outperforms the larger non-distilled Qwen2.5-1.5B and Phi-1.5 models on Urban-Text by 1.7% and 1.9%, respectively. Consistent improvements are observed for Flan-T5 and TinyLlama, whose Urban-Text F1 scores increase by 1.8% and 1.7% with distillation. Although the distilled Qwen2.5-1.5B achieves a relative Urban-Text F1 advantage of 0.9% over the distilled 0.5B model, the latter requires 67.3% less inference time in this text evaluation. Notably, the improvements extend to downstream trajectory prediction on nuScenes. As shown in Table 10, CoT-based distillation reduces minADE<sub>5</sub> by 12.7% for Qwen2.5-0.5B relative to its LoRA-only counterpart. Corresponding reductions are 7.7% for Flan-T5, 10.2% for both TinyLlama and Phi-1.5, and 11.2% for Qwen2.5-1.5B. Furthermore, the distilled 0.5B model achieves 6.4% and 7.9% lower minADE<sub>5</sub> than the larger non-distilled Qwen2.5-1.5B and Phi-1.5 models, respectively. Together with the text-evaluation results, these comparisons support the effectiveness of the proposed distillation strategy across the evaluated model families.

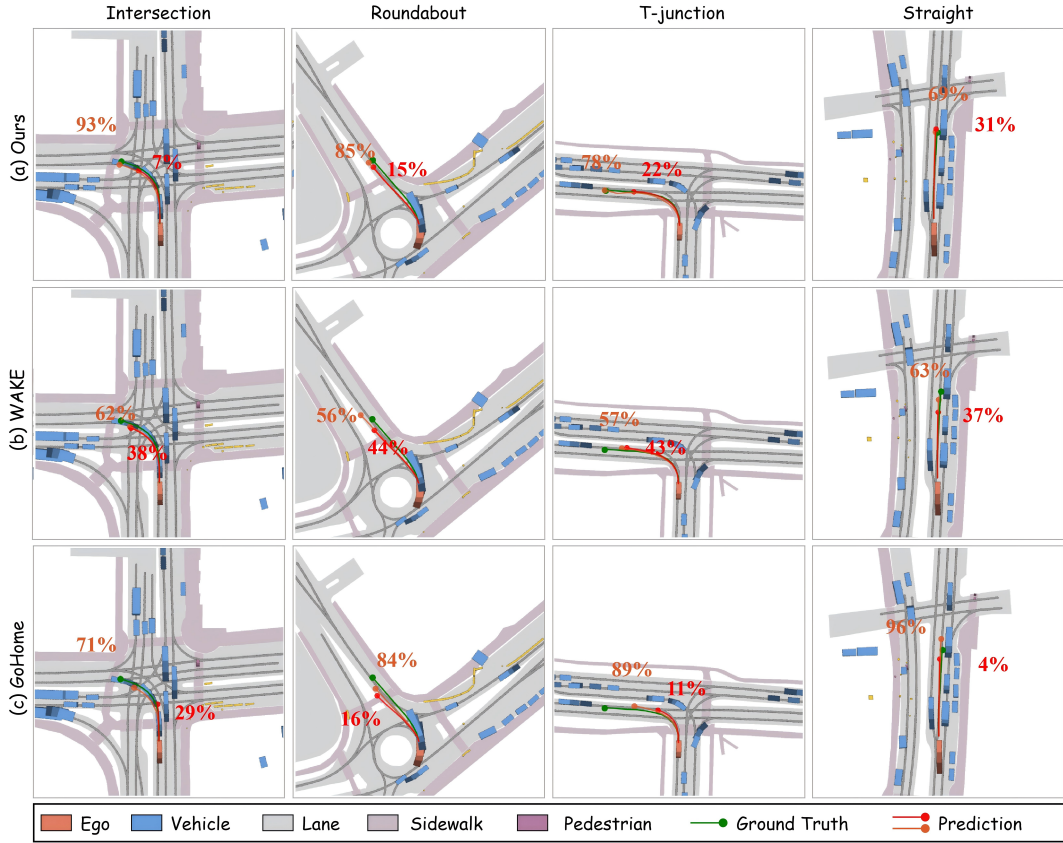
In addition, Table 10 reports a prediction latency of 237 ms for the distilled Qwen2.5-0.5B configuration, representing reductions of 50.2%, 58.6%, and 59.6% relative to the distilled TinyLlama, Phi-1.5, and Qwen2.5-1.5B configurations, respectively. The distilled Qwen2.5-1.5B provides a further 5.1% reduction in minADE<sub>5</sub> relative to the distilled 0.5B model, accompanied by a higher inference cost. Considering both its improvements over LoRA-only baselines and its lower latency relative to larger distilled models, we select distilled Qwen2.5-0.5B as the compact semantic backbone of LLM-TP<sup>2</sup>, offering a favorable accuracy-efficiency balance among the evaluated configurations.

## 4.5. Qualitative Results

### 4.5.1. Qualitative Results for Trajectory Prediction

To further examine how the proposed framework behaves in challenging urban scenes, we provide a qualitative comparison with two recent baselines, WAKE and GoHome, on representative nuScenes examples. As illustrated in Fig. 4, the comparison spans four typical urban topologies, namely intersections, roundabouts, T-junctions, and straight road segments. The visualization highlights not only the geometric accuracy of the predicted trajectories, but also the quality of multimodal intent allocation under different map constraints and interaction contexts.

**Intersection Scene.** As shown in Fig. 4 (Intersection), the ego agent is approaching a highly structured intersection where multiple maneuver choices coexist. In this case, LLM-TP<sup>2</sup> assigns the dominant probability mass to the turning trajectory that closely matches the ground truth, while suppressing implausible alternatives. By contrast, WAKE



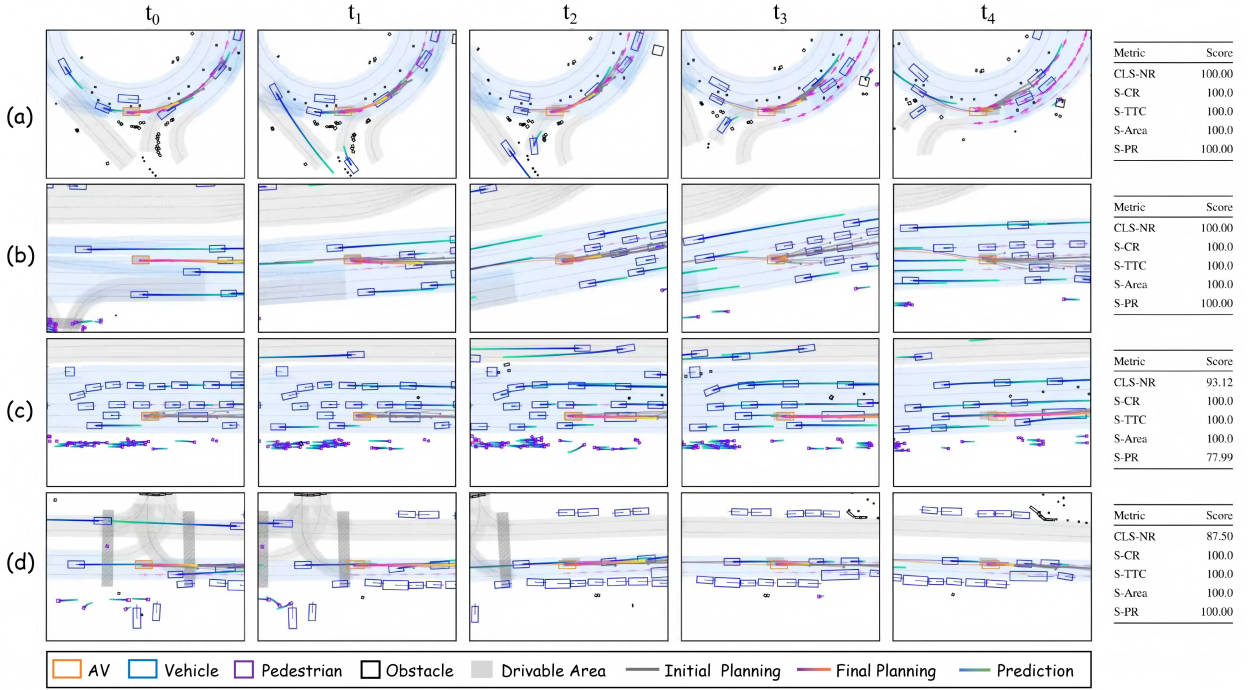
**Figure 4:** Qualitative evaluation of multimodal trajectory prediction across diverse urban scenes from the nuScenes dataset. The rows correspond to the evaluated models: the proposed LLM-TP<sup>2</sup> framework, followed by the WAKE [78] and GoHome [7] baselines. The columns illustrate four distinct interactive scenarios: a complex intersection, a curved roundabout, a T junction, and a straight road segment. In each panel, the ground truth future trajectory is highlighted in green, while the predicted multimodal candidates are depicted in red and orange alongside their inferred confidence probabilities. For clarity, the number of predicted trajectories is set to 2. Compared with the baselines, LLM-TP<sup>2</sup> produces the predictions that are more topologically consistent with the road layout and more spatially aligned with the ground truth.

produces a more ambiguous mode distribution, and GoHome exhibits a larger spatial deviation in the predicted turning path. This comparison suggests that the proposed framework can more effectively combine historical motion cues with lane-level topology, thereby reducing multimodal uncertainty at decision-critical branching regions.

**Roundabout Scene.** Fig. 4 (Roundabout) presents a roundabout scenario with curved lane geometry and continuous heading changes. Such scenes are particularly challenging because the feasible future motion is strongly constrained by the circular road structure, while the local curvature may be difficult to extrapolate from short motion history alone. In this example, LLM-TP<sup>2</sup> generates a predicted trajectory that remains tightly aligned with the roundabout lane and closely follows the ground truth. In comparison, WAKE shows noticeably less stable curvature adherence, and GoHome introduces a larger offset between the predicted path and the actual vehicle motion. This result indicates that the proposed framework better captures non-linear topological constraints in underrepresented road structures.

**T-junction Scene.** As shown in Fig. 4 (T-junction), the ego vehicle is required to resolve a clear topological choice at a junction with limited outgoing branches. In this case, LLM-TP<sup>2</sup> produces an accurate multimodal prediction, with the dominant mode closely matching the ground-truth turning behavior. Although both baselines also identify the general maneuver direction, their predicted trajectories either deviate more from the lane centerline or allocate higher confidence to less accurate alternatives. This example shows that the proposed framework can more reliably infer intent in constrained scenarios, where precise alignment between map structure and maneuver semantics is essential.





**Figure 5:** Qualitative closed-loop evaluation results of our proposed LLM-TP<sup>2</sup> on four representative nuPlan scenarios. The time sequence from  $t_0$  to  $t_4$  illustrates the adaptive behavior of the AV (orange box) in complex traffic environments. Panels (a)–(d) showcase: (a) roundabout navigation, (b) high-speed lane changing, (c) multi-agent interactions in dense traffic, and (d) car-following integrated with pedestrian awareness. In each case, the initial plan (gray) is progressively refined into the final plan (purple) under the guidance of surrounding-agent prediction (blue) and map constraints, yielding safe and feasible motion in complex traffic environments. Accompanying tables on the right provide the quantitative evaluation scores for key metrics, including safety (S-CR, S-TTC, S-Area) and planning progress (CLS-NR, S-PR), demonstrating the superior capability of LLM-TP<sup>2</sup> to generate safe and efficient movements in challenging urban scenarios.

**Straight-Road Scene.** Fig. 4 (Straight) illustrates a seemingly simple straight-road scenario. Although the road geometry is less ambiguous than in intersections or roundabouts, accurate prediction still requires the correct estimation of longitudinal motion and local interaction context. In this case, LLM-TP<sup>2</sup> produces a future trajectory that better overlaps with the ground truth and yields a more reasonable confidence distribution over candidate modes. By comparison, both WAKE and GoHome show larger discrepancies in the predicted forward evolution. This observation suggests that the advantage of LLM-TP<sup>2</sup> is not limited to highly non-linear scenes, but also extends to speed-sensitive road segments where longitudinal behavior must be estimated precisely.

Overall, these examples show that LLM-TP<sup>2</sup> consistently generates predictions that are more topologically coherent, spatially accurate, and behaviorally plausible across diverse urban layouts.

#### 4.5.2. Qualitative Results for Closed-Loop Planning

To further evaluate the closed-loop performance of LLM-TP<sup>2</sup>, we select four representative and challenging scenarios from the nuPlan benchmark for qualitative analysis. As illustrated in Fig. 5, the visualization captures the AV's (orange box) trajectory evolution from  $t_0$  to  $t_4$ , where the final planned trajectory (purple) adaptively adjusts based on the predicted paths of surrounding agents (blue). The specific visual analysis is as follows:

**Roundabout Scene.** As shown in Fig. 5 (a), the AV enters and traverses a roundabout under dense surrounding traffic. From  $t_0$  to  $t_4$ , multiple vehicles approach and circulate within the roundabout, leading to strong interaction uncertainty and limited maneuvering space. In this case, the AV follows a smooth and collision-free trajectory while maintaining compatibility with the circulating traffic flow, rather than performing overly aggressive insertion or unnecessary braking. The final plan remains well aligned with the predicted motions of surrounding agents. This



behavior is consistent with the perfect closed-loop scores reported on the right, where all metrics, including CLS-NR, S-CR, S-TTC, S-Area, and S-PR, reach 100.0, indicating that LLM-TP<sup>2</sup> can effectively handle non-linear road geometries without compromising safety or efficiency.

**Lane-Change Scene.** Fig. 5 (b) showcases a strategic lane-merging maneuver on a high-speed road. Across  $t_0$  to  $t_3$ , the AV avoids abrupt steering actions and instead executes a gradual lane transition that preserves both lateral safety margins and longitudinal consistency with neighboring traffic. As the scene progresses to  $t_4$ , the final planning trajectory (purple) smoothly transitions into the target lane while maintaining a safe longitudinal buffer from both preceding and succeeding agents. Notably, all closed-loop metrics remain at 100.0, indicating that the generated trajectory preserves both safety and route consistency throughout the maneuver.

**Multi-Agent Interaction Scene.** Fig. 5 (c) shows a dense multi-agent traffic scene with numerous surrounding vehicles and complex interaction patterns, where the AV must negotiate space with numerous interacting agents. Despite the crowded environment and multiple overlapping motion tendencies, the AV is able to maintain a feasible trajectory within the drivable corridor and avoid unsafe acceleration or frequent lane-cutting. While the progress score (S-PR) is recorded at 77.99 due to the inherent constraints of the dense traffic, the safety-related metrics (S-CR, S-TTC, and S-Area) remain at a maximum of 100.0. This underscores the risk-aware nature of LLM-TP<sup>2</sup>, which prioritizes collision avoidance and stable interaction over aggressive progress in high-interaction scenarios.

**Car-Following Scene.** As illustrated in Fig. 5 (d), the AV performs longitudinal planning in a structured car-following scenario with both vehicles and vulnerable road users (VRUs) in the vicinity. From  $t_0$  to  $t_4$ , the model successfully monitors the VRUs near the crosswalk areas while maintaining a consistent distance from the lead vehicle. Rather than producing overly conservative stopping behavior or aggressive closing maneuvers, the planner generates a balanced trajectory that is both safe and efficient, achieving high scores in both CLS-NR (87.50) and safety metrics (100.0). These results demonstrate that the proposed framework possesses robust cross-modal understanding and real-time adaptability, ensuring safe navigation in structured urban scenes with mixed traffic participants.

Overall, these examples demonstrate that the LLM-TP<sup>2</sup> framework consistently refines the initial proposal into a safer, more context-aware final trajectory during closed-loop execution. The accompanying scenario-wise metrics corroborate the visual results, showing that the proposed framework maintains strong safety performance across diverse driving situations and adapts its progressiveness to scene complexity.

#### 4.5.3. Qualitative Results for Ablation Models of Prediction

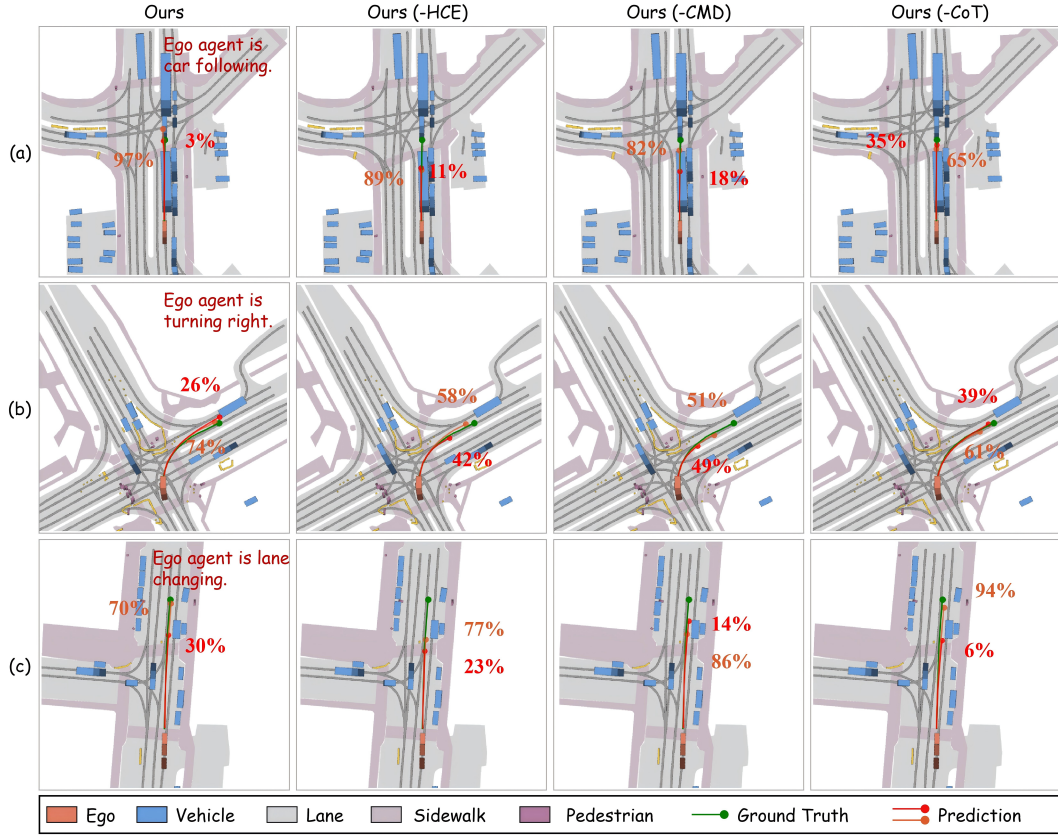
To further demonstrate the contribution of each proposed module, we present a qualitative ablation analysis across three representative driving scenarios: car-following, right-turning, and lane-changing. As shown in Fig. 6, the comparison evaluates the full LLM-TP<sup>2</sup> model against three ablated variants: without the Hierarchical Contextual Encoder (-HCE), without the cross-modal Decoder (-CMD), and without Chain-of-Thought distillation (-CoT).

**Car-Following Scene.** As illustrated in Fig. 6(a), the full model precisely captures the longitudinal speed profile, maintaining a trajectory tightly aligned with the ground truth. While omitting the CoT distillation slightly degrades precision, removing the HCE or CMD results in significant positional deviations. This highlights the necessity of robust semantic scene understanding and cross-modal grounding for maintaining accurate following behaviors.

**Right-Turning Scene.** Fig. 6 (b) presents a challenging right-turning scenario characterized by clear topological branching. The full model successfully navigates the intersection, maintaining high geometric fidelity to the ground-truth path. Among the ablated variants, the model without CoT distillation (-CoT) preserves a reasonable turning profile but lacks fine-grained curvature precision. Conversely, omitting either the HCE or CMD results in pronounced trajectory deviations, revealing a critical deficit in topological reasoning and spatial grounding. This scenario demonstrates that robust prediction at complex branches fundamentally relies on semantic scene comprehension and cross-modal grounding, with CoT distillation providing essential trajectory refinement.

**Lane-Changing Scene.** As shown in Fig. 6 (c), the full framework accurately infers the lateral transition intent. Removing the HCE severely impairs the model's ability to recognize interaction-driven maneuver changes, whereas omitting the CMD degrades the geometric smoothness of the lateral shift.

Overall, these qualitative visualizations perfectly corroborate the quantitative findings in Table 8. The HCE is crucial for abstracting road topology and agent intent; the CMD provides the indispensable geometric grounding to translate semantic cues into the coordinate space; and the CoT distillation explicitly refines interaction reasoning. The synergy of these components ensures robust and precise trajectory generation across diverse urban dynamics.



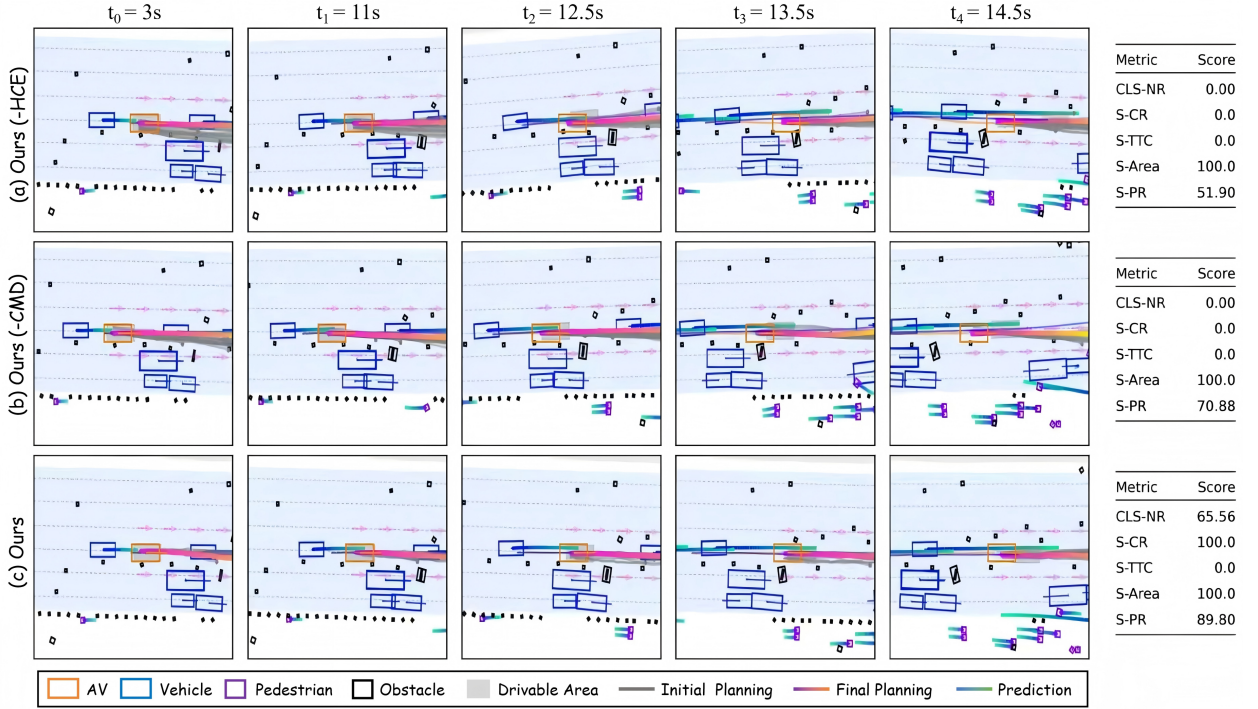
**Figure 6:** Qualitative ablation analysis of trajectory prediction across nuScenes dataset. The columns sequentially display the full LLM-TP<sup>2</sup> framework, the model omitting the Hierarchical Contextual Encoder (-HCE), the model omitting the cross-modal Decoder (-CMD), and the model omitting the CoT-based distillation (-CoT). The rows depict distinct interactive maneuvers: (a) longitudinal car following, (b) right turning, and (c) lateral lane changing. In each panel, the ground truth trajectory is visualized in green, accompanied by the predicted multimodal candidate trajectories and their associated confidence probabilities. LLM-TP<sup>2</sup> consistently generates predictions that are better aligned with the road topology and future vehicle motion, whereas the ablated variants exhibit degraded intent recognition and weaker spatial grounding.

#### 4.5.4. Qualitative Results for Ablation Methods of Planning

To intuitively evaluate the contribution of each architectural component, we conduct a comparative visual analysis using representative corner cases from the nuPlan dataset. As illustrated in Fig. 7, the scenario involves an ego-vehicle navigating through a high-density environment with multiple static obstacles and dynamic traffic agents.

**Impact of Semantic Reasoning (-HCE).** As shown in Fig. 7 (a), the model operating without the Hierarchical Contextual Encoder (-HCE) exhibits a total failure in strategic planning. Without the LLM-driven semantic reasoning module, the model can only rely on geometric motion cues and cannot thus infer the higher-level causal relationships and interaction intent within the scene. Consequently, the AV fails to perceive the high-level risks associated with the surrounding agents, leading to a zero score in both planning progress (CLS-NR) and safety (S-CR).

**Impact of cross-modal Fusion (-CMD).** Fig. 7 (b) shows the performance of the variant lacking the cross-modal Decoder (-CMD). Ours (-CMD) variant performs better than Ours (-HCE), but still shows clear deficiencies. Although it may possess certain semantic insights from the encoder, the absence of an effective attention-based bridge between linguistic priors and spatial trajectory features prevents the realization of safe maneuvers. Consequently, the planner produces a feasible but less coordinated rollout, with weaker consistency between semantic understanding and motion execution. While it achieves a moderate progress score (S-PR = 70.88), it ultimately fails all critical safety benchmarks with an S-CR of 0.0. This result establishes that even if a model possesses semantic understanding, it cannot execute safe maneuvers without a robust mechanism to project those cognitive priors into the final trajectory planning space.



**Figure 7:** Qualitative closed-loop comparison of LLM-TP<sup>2</sup> and its ablated method on a representative nuPlan scene. Panels (a)-(c) show Ours (-HCE), Ours (-CMD), and the full model, respectively. The “-HCE” variant removes the Hierarchical Contextual Encoder, whereas “-CMD” removes the cross-modal attention in the decoder. Across  $t_0$ - $t_4$ , the full model (Ours) generates a smoother and more context-aware trajectory that better matches the surrounding traffic evolution. The metrics on the right further support this observation: the complete model achieves higher CLS-NR (65.56) and S-PR (89.80) than Ours (-HCE) (0.00, 51.90) and Ours (-CMD) (0.00, 70.88), while preserving perfect S-CR and S-Area scores.

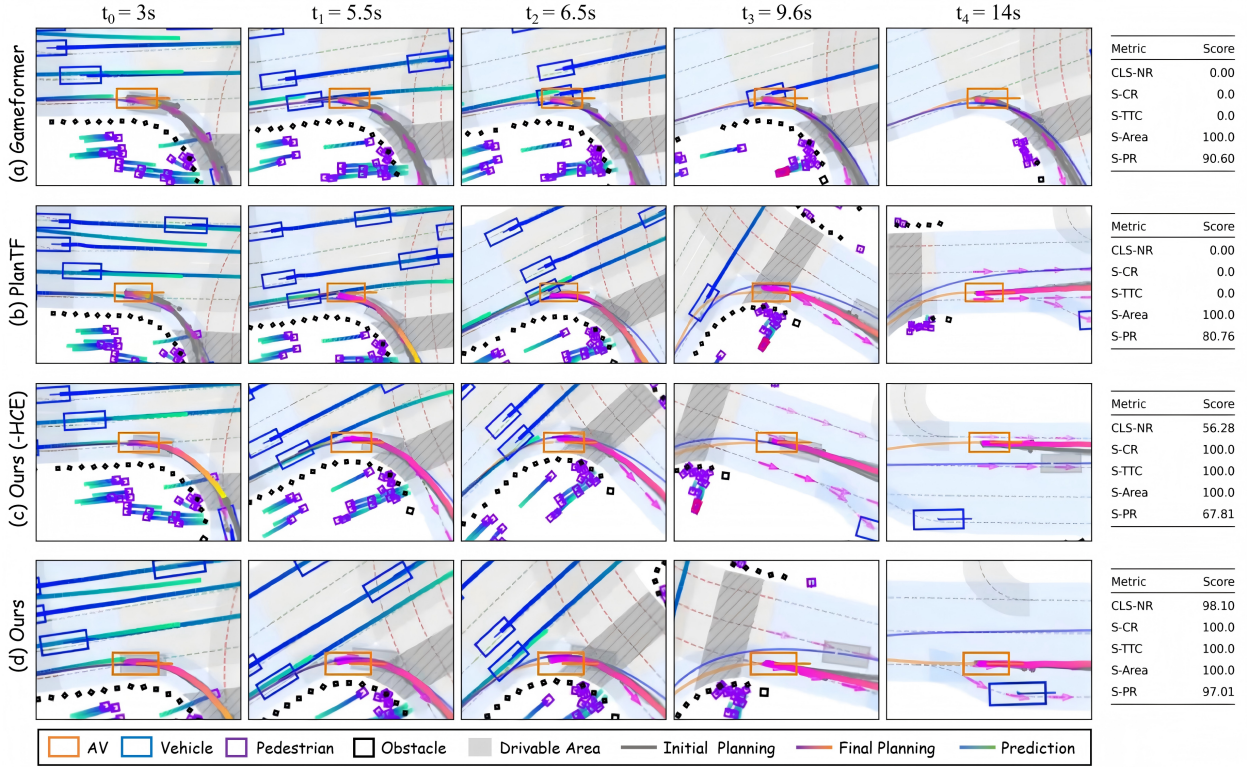
**Performance of the Full Model (Ours).** In contrast, our complete LLM-TP<sup>2</sup> model shown in Fig. 7 (c) achieves the optimal balance between reasoning depth and execution precision. By combining hierarchical causal reasoning with precise cross-modal alignment, the full LLM-TP<sup>2</sup> accurately evaluates environmental constraints and generates a smoother planning trajectory that better aligns with the predicted motions of surrounding agents. It maintains safer interaction patterns while also preserving route progress in the dense traffic scene. This qualitative improvement is further supported by the metrics on the right, where CLS-NR rises to 65.56, and S-PR reaches 89.80, while S-CR and S-Area remain at 100.0. Overall, these qualitative comparisons highlight the complementary roles of HCE and CMD in closed-loop planning. HCE provides high-level scene understanding and semantic intent reasoning, while CMD effectively injects this semantic information into geometric trajectory generation. Removing either component weakens the planner, while their joint use yields the most robust and progressive behavior in interactive scenes.

#### 4.6. Case Study

As shown in Fig. 8, we present a closed-loop case study on a highly interactive curved-road scene from nuPlan, where the ego agent must negotiate surrounding vehicles and vulnerable road users (VRUs) within a narrow drivable corridor. This example compares the hybrid-learning baseline GameFormer, the pure deep learning baseline PlanTF, the ablated variant Ours (-HCE), and the full LLM-TP<sup>2</sup> model (Ours).

As shown in Fig. 8 (a) and (b), both GameFormer and PlanTF produce unsafe rollouts during closed-loop execution. Although their trajectories initially follow the road geometry, both exhibit hesitation during the turn and fail to maintain safe interaction margins as the scene evolves, eventually resulting in collisions. This failure is reflected in the quantitative results, where both methods obtain zero scores on CLS-NR, S-CR, and S-TTC. While their S-Area remains 100.0, indicating that the rollouts stay within the drivable area, this alone is insufficient for safe closed-loop driving in dense interactive scenes. Their relatively high S-PR scores further suggest that these methods preserve forward progress at the expense of interaction safety. Moreover, the ablated model, Ours (-HCE), shown in Fig. 8(c),





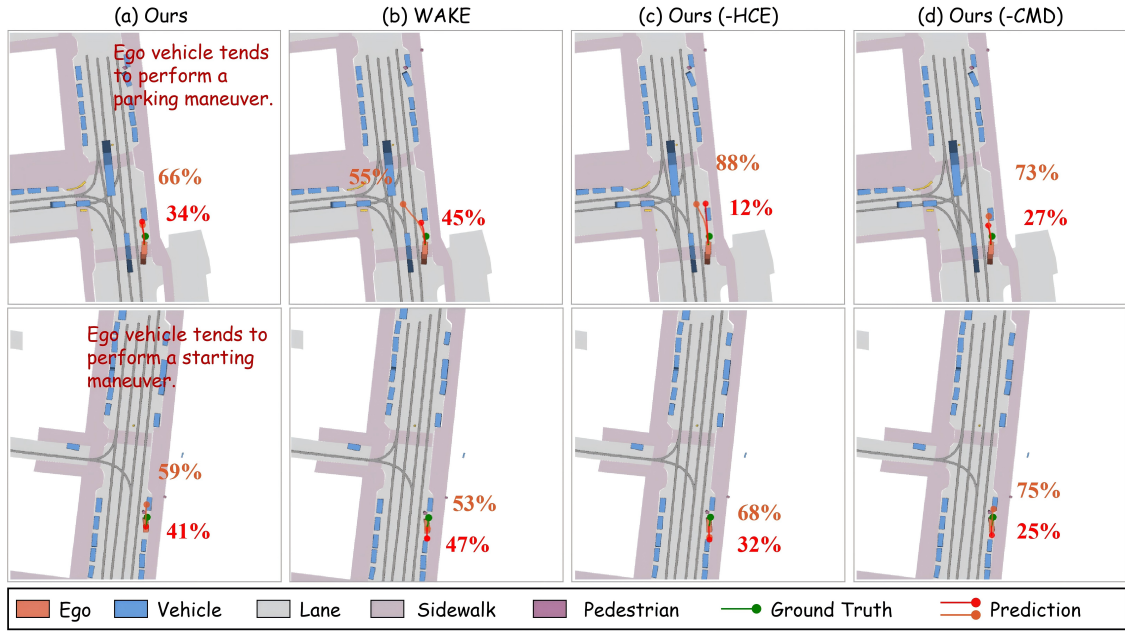
**Figure 8:** Qualitative comparison of closed-loop planning performance between LLM-TP<sup>2</sup> and SOTA baselines in a complex right-turn intersection scenario within the nuPlan benchmark. Panels (a) to (d) illustrate the behavior of: (a) Gameformer, (b) PlanTF, (c) Ours without the Hierarchical Contextual Encoder (-HCE), and (d) our full LLM-TP<sup>2</sup> framework. The time sequence from  $t_0 = 3s$  to  $t_4 = 14s$  visualizes the AV (orange box) navigating through high-density pedestrian interactions and surrounding traffic. Compared with the hybrid learning-based Gameformer and deep learning-based PlanTF, the proposed model generates a smoother, more interaction-consistent trajectory that more closely follows the drivable corridor and surrounding traffic flow. The metrics on the right further confirm this advantage: our proposed LLM-TP<sup>2</sup> model achieves near-perfect performance in both safety and planning progress (S-CR = 100.0, CLS-NR = 98.10).

performs substantially better in safety. It avoids collisions and achieves a perfect S-CR of 100.0. However, without LM/LLM-based semantic reasoning, it becomes more conservative and less route-consistent, as reflected by its lower CLS-NR of 56.28 and S-PR of 67.81. This indicates that geometric features alone can provide basic safety, but are insufficient for maintaining stable and efficient progress in complex interactive driving scenarios.

As shown in Fig. 8 (d), the full model achieves the best overall balance between safety, progress, and route consistency. Its rollout remains smooth and collision-free while adapting coherently to the surrounding traffic. Quantitatively, it attains the highest CLS-NR of 98.10 and S-PR of 97.01, while preserving perfect scores on S-CR, S-TTC, and S-Area. Compared with Ours (-HCE), the large gains in CLS-NR and S-PR show that the hierarchical contextual encoder provides critical scene-level reasoning that improves long-horizon route execution and planning efficiency. Compared with GameFormer and PlanTF, our model not only avoids safety failures but also achieves stronger progress without sacrificing robustness. Overall, this case study highlights two main observations. First, geometric feasibility alone is not sufficient for safe closed-loop driving, since limited reasoning over complex interactions can still lead to unsafe behavior. Second, semantic reasoning improves not only safety but also the coherence of closed-loop planning. By incorporating the HCE, LLM-TP<sup>2</sup> better interprets scene context and anticipates interaction evolution, producing trajectories that are safer, more progressive, and more route-consistent.

#### 4.7. Failure Case Analysis and Discussion

To provide a comprehensive assessment of the proposed framework, this section conducts a rigorous diagnostic analysis of scenarios where LLM-TP<sup>2</sup> and existing SOTA methods encounter performance degradation. By examining

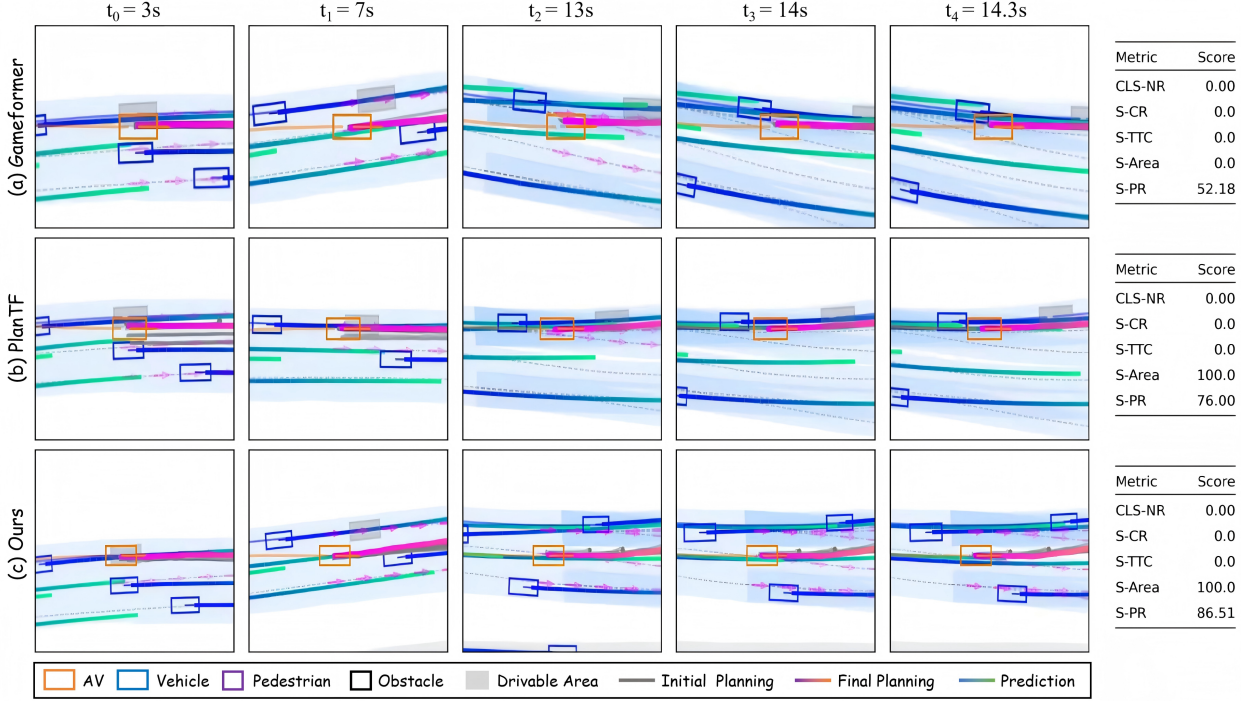


**Figure 9:** Failure-case comparison of trajectory prediction on challenging nuScenes scenarios. The columns correspond to different models: (a) the proposed LLM-TP<sup>2</sup>, (b) the SOTA baseline WAKE [78], (c) our model omitting the Hierarchical Contextual Encoder (-HCE), and (d) our model omitting the Cross-modal Decoder (-CMD). The top row illustrates a complex parking entry maneuver, while the bottom row depicts a starting transition situation. Although all methods degrade in these cases, LLM-TP<sup>2</sup> produces predictions that remain closer to the ground truth, especially in the starting-transition scenario, where semantic reasoning helps mitigate the ambiguity caused by misleading motion history.

these corner cases in both open-loop forecasting and closed-loop planning, we aim to uncover the fundamental challenges posed by the long-tail distribution of traffic behaviors.

**Failure Case Analysis of Prediction.** Fig. 9 presents two representative failure cases, namely parking entry and starting transition, to illustrate the performance boundary of the proposed framework in long-tail scenarios. These cases mainly arise from two sources: insufficient training coverage for rare maneuvers and strong conflict between historical motion cues and future intent. In the parking-entry case shown in the first row, all models fail to accurately capture the parking behavior and instead produce predictions that are closer to an overtaking pattern. This failure is likely caused by the limited number of similar samples in the training data, which makes such rare maneuvers difficult to model reliably. Although LLM-TP<sup>2</sup> still produces a relatively more plausible prediction, it remains affected by the same data sparsity issue. The second row shows a starting-transition case, where the future motion is inconsistent with the historical stationary state. In this scenario, methods without the HCE module are more easily dominated by motion inertia and therefore tend to predict continued stationary or straight-motion behavior. By comparison, LLM-TP<sup>2</sup> better captures the latent starting intent and produces a prediction that is more consistent with the ground truth. Overall, these examples suggest that semantic reasoning improves robustness when future intent cannot be inferred reliably from motion history alone. However, its benefit still depends on sufficient data coverage. In extremely rare driving scenarios, even strong semantic priors may not fully prevent prediction errors.

**Failure Case Analysis of Planning.** While LLM-TP<sup>2</sup> achieves strong overall performance, certain highly interactive scenarios remain challenging for all methods. Fig. 10 shows one such case, where severely limited maneuvering space and rapidly changing surrounding-agent motions create a tightly coupled closed-loop environment. In this case, the AV must progress through a narrow drivable corridor while continuously adapting to nearby vehicles whose future behaviors strongly influence one another. As a result, all compared approaches, including GameFormer, PlanTF, and LLM-TP<sup>2</sup>, fail to maintain a collision-free trajectory, yielding zero scores on both S-CR and CLS-NR. A likely cause of failure in this scenario is the combination of severely constrained maneuvering space and rapidly evolving multi-agent coupling. The surrounding traffic configuration leaves only a narrow feasible corridor for the AV, while the relative motions of nearby agents continuously alter this corridor over time. In such settings, closed-loop planning depends



**Figure 10:** Failure-case comparison in closed-loop evaluation on a challenging nuPlan scenario. Panels (a)-(c) correspond to GameFormer, PlanTF, and LLM-TP<sup>2</sup> (Ours), respectively, while columns  $t_0$ - $t_4$  illustrate the temporal evolution of the closed-loop rollout. In this extremely challenging high-interaction environment, all models encounter safety violations and fail to meet the strict collision-free criteria, resulting in zero scores for CLS-NR, S-CR, and S-TTC. Nevertheless, compared with the baseline methods, our proposed LLM-TP<sup>2</sup> still produces a smoother and more progressive trajectory, achieving the highest S-PR score (86.51), versus 52.18 for the GameFormer and 76.00 for the PlanTF.

not only on maintaining local geometric feasibility but also on accurately anticipating the evolution of the surrounding agents. As shown in Fig. 10 (a) and (b), the baselines struggle to balance these requirements, leading to unstable rollouts or eventual collisions with neighboring vehicles. More broadly, when the predicted future behaviors of surrounding agents become even slightly misaligned with the actual closed-loop evolution, the planner may follow a trajectory that is locally plausible but no longer interaction-safe after several steps. This indicates that semantic or contextual priors can provide valuable guidance, but may still be insufficient when interaction uncertainty is high, and the available safety margin is extremely limited. The comparison with GameFormer and PlanTF further highlights the difficulty of current prediction-and-planning frameworks in such long-tail interactive scenes. Both baselines fail on the core safety metrics, with S-CR and S-TTC dropping to 0.00. GameFormer also yields an S-Area score of 0.0, indicating that its rollout eventually violates drivable-area constraints, whereas PlanTF preserves S-Area of 100.0 but still fails to maintain interaction safety. This difference suggests that geometric feasibility alone does not necessarily ensure robust closed-loop behavior under strong interaction complexity. By comparison, LLM-TP<sup>2</sup> achieves the highest S-PR score of 86.51, outperforming GameFormer (52.18) and PlanTF (76.00), and yields a relatively smoother rollout throughout the scene. Although the outcome remains unsuccessful, this suggests that the proposed framework preserves stronger motion stability and route progress under failure conditions. This behavior is likely attributable to the semantic reasoning module, which helps the planner exploit contextual cues and maintain a more coherent trajectory under uncertainty. At the same time, this case also indicates that semantic enhancement alone may be insufficient in extremely constrained interactive settings, where rapid contraction of the feasible set can still cause errors to accumulate.

The collective observations from these failure modes highlight a critical hierarchy of challenges in autonomous driving: while semantic reasoning effectively mitigates logical errors in intent recognition, it cannot entirely circumvent the physical constraints imposed by high-density interactions. The transition from logical intent to safe geometric execution remains a sensitive bottleneck. Importantly, these findings suggest several directions for future research.



First, the gap between high-level semantic reasoning and low-level geometric execution could be further reduced by introducing stronger safety-aware mechanisms within the Cross-modal decoding stage, such as counterfactual reasoning or additional formal safety constraints. Second, improved modeling of rapid behavioral change points may enable better real-time uncertainty calibration in highly interactive scenes. Overall, while LLM-TP<sup>2</sup> effectively exploits semantic priors to push the boundaries of current performance, further research is required to convert these reasoning gains into deterministic safety guarantees in rare and safety-critical driving conditions.

## 5. Conclusion

This paper introduces LLM-TP<sup>2</sup>, a unified framework for joint trajectory prediction and planning that integrates high-level semantic reasoning with real-time decision-making. To pursue the trade-off between computational efficiency and reasoning depth, we propose a Hierarchical Contextual Encoder that adaptively triggers LLM-based semantic reasoning by monitoring environmental uncertainty. This mechanism allows the model to allocate intensive reasoning resources only during complex driving scenarios, ensuring both operational efficiency and safety. By employing a CoT-based latent distillation paradigm, we transfer complex contextual priors into a compact 0.5B student model, while a Cross-modal decoder aligns linguistic insights with geometric trajectory features. Extensive evaluations on real-world datasets and closed-loop benchmarks demonstrate that LLM-TP<sup>2</sup> significantly outperforms SOTA baselines in terms of trajectory prediction and planning. Despite these improvements, our failure analysis reveals a remaining gap between high-level semantic intent and low-level geometric execution during dense multi-agent interactions. Future research will focus on narrowing this discrepancy through continuous learning and refined semantic modeling to further improve the reliability of AV decision-making in safety-critical environments. Moreover, we will also investigate systematic annotation verification and assess the sensitivity of semantic distillation to annotation noise.

## CRedit authorship contribution statement

**Haicheng Liao:** Conceptualization, Methodology, Experiment, Writing. **Chengyue Wang:** Methodology, Experiment, Writing. **Hanlin Kong:** Experiment. **Zhenning Li:** Conceptualization, Methodology, Funding. **Zhengbing He:** Methodology. **Zhiyong Cui:** Methodology. **Guohui Zhang:** Methodology, Review. **Chengzhong Xu:** Funding.

## Acknowledgements

This work was supported by the Science and Technology Development Fund of Macau [0007/2025/RIC, 0122/2024/RIB2, 0215/2024/AGJ, 0074/2025/AMJ, 001/2024/SKL, 0002/2025/EQP], the Research Services and Knowledge Transfer Office, University of Macau [SRG2023-00037-IOTSC, MYRG-GRG2024-00284-IOTSC], the Shenzhen-Hong Kong-Macau Science and Technology Program Category C [SGDX20230821095159012], the Science and Technology Planning Project of Guangdong [2025A0505010016], National Natural Science Foundation of China [52572354], the State Key Lab of Intelligent Transportation System [2024-B001], and the Jiangsu Provincial Science and Technology Program [BZ2024055]. Please contact Dr. Zhenning Li (zhenningli@um.edu.mo) and Dr. Chengzhong Xu (czxu@um.edu.mo) for correspondence.

## References

- [1] Z. Li, Z. Cui, H. Liao, J. Ash, G. Zhang, C. Xu, Y. Wang, Steering the future: Redefining intelligent transportation systems with foundation models, *Chain 1* (2024) 46–53.
- [2] H. Liao, Z. Li, K. Zhu, K. Li, C. Xu, Sa-tp<sup>2</sup>: A safety-aware trajectory prediction and planning model for autonomous driving, *IEEE Transactions on Robotics* 41 (2025) 5267–5286.
- [3] M. Treiber, A. Hennecke, D. Helbing, Congested traffic states in empirical observations and microscopic simulations, *Physical review E* 62 (2000) 1805.
- [4] D. Dauner, M. Hallgarten, A. Geiger, K. Chitta, Parting with misconceptions about learning-based vehicle motion planning, in: *CoRL*, 2023, pp. 1268–1281.
- [5] R. Wang, S. Wang, H. Yan, X. Wang, Wisp: wave superposition inspired pooling for dynamic interactions-aware trajectory prediction, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 2023, pp. 4685–4692.
- [6] Y. Ren, Z. Lan, L. Liu, H. Yu, Emsin: Enhanced multi-stream interaction network for vehicle trajectory prediction, *IEEE Transactions on Fuzzy Systems* (2024).
- [7] T. Gilles, S. Sabatini, D. Tsishkou, B. Stanciulescu, F. Moutarde, Gohome: Graph-oriented heatmap output for future motion estimation, in: *ICRA*, IEEE, 2022, pp. 9107–9114.

- [8] J. Cheng, Y. Chen, X. Mei, B. Yang, B. Li, M. Liu, Rethinking imitation-based planners for autonomous driving, in: ICRA, IEEE, 2024, pp. 14123–14130.
- [9] Z. Fu, K. Jiang, T. Wen, H. Z. Lim, J. Huang, M. Yang, H. Gao, Z. Zhong, D. Yang, C2f-planner: Interaction-aware coarse-to-fine planning for autonomous vehicles, IEEE Robotics and Automation Letters (2025).
- [10] H. Caesar, J. Kabzan, K. S. Tan, W. K. Fong, E. Wolff, A. Lang, L. Fletcher, O. Beijbom, S. Omari, nuplan: A closed-loop ml-based planning benchmark for autonomous vehicles, arXiv preprint arXiv:2106.11810 (2021).
- [11] J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang, B. Hui, L. Ji, M. Li, J. Lin, R. Lin, D. Liu, G. Liu, C. Lu, K. Lu, J. Ma, R. Men, X. Ren, X. Ren, C. Tan, S. Tan, J. Tu, P. Wang, S. Wang, W. Wang, S. Wu, B. Xu, J. Xu, A. Yang, H. Yang, J. Yang, S. Yang, Y. Yao, B. Yu, H. Yuan, Z. Yuan, J. Zhang, X. Zhang, Y. Zhang, Z. Zhang, C. Zhou, J. Zhou, X. Zhou, T. Zhu, Qwen technical report, arXiv preprint arXiv:2309.16609 (2023).
- [12] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al., Llama: Open and efficient foundation language models, arXiv preprint arXiv:2302.13971 (2023).
- [13] Z. Lan, L. Liu, B. Fan, Y. Lv, Y. Ren, Z. Cui, Traj-llm: A new exploration for empowering trajectory prediction with pre-trained large language models, IEEE Transactions on Intelligent Vehicles (2024).
- [14] H. Liao, B. Wang, J. Yang, C. Wang, Z. He, G. Zhang, C. Xu, Z. Li, Addressing corner cases in autonomous driving: A world model-based approach with mixture of experts and llms, Transportation Research Part C: Emerging Technologies 183 (2026) 105456.
- [15] I. Bae, J. Lee, H.-G. Jeon, Can language beat numerical regression? language-based multimodal trajectory prediction, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 753–766.
- [16] Z. Yang, X. Jia, H. Li, J. Yan, Llm4drive: A survey of large language models for autonomous driving, 2023. arXiv:2311.01043.
- [17] H. Liao, B. Kong, B. Wang, C. Wang, W. Ye, Z. He, C. Xu, Z. Li, Cot-drive: Efficient motion forecasting for autonomous driving with llms and chain-of-thought prompting, IEEE Transactions on Artificial Intelligence (2025).
- [18] H. Atici-Ulusu, O. Taskapilioglu, T. Gunduz, A neuroergonomics approach to investigate the mental workload of drivers in real driving settings, Transportation research part F: traffic psychology and behaviour 103 (2024) 177–189.
- [19] K. Friston, The free-energy principle: a unified brain theory?, Nature reviews neuroscience 11 (2010) 127–138.
- [20] Y. Hu, J. Yang, L. Chen, K. Li, C. Sima, X. Zhu, S. Chai, S. Du, T. Lin, W. Wang, et al., Planning-oriented autonomous driving, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 17853–17862.
- [21] X. Weng, B. Ivanovic, Y. Wang, Y. Wang, M. Pavone, Para-drive: Parallelized architecture for real-time autonomous driving, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 15449–15458.
- [22] B. Zhang, N. Song, X. Jin, L. Zhang, Bridging past and future: End-to-end autonomous driving with historical prediction and planning, in: Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 6854–6863.
- [23] I. Bae, H.-G. Jeon, A set of control points conditioned pedestrian trajectory prediction, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 37, 2023, pp. 6155–6165.
- [24] M. Zare, P. M. Kebria, A. Khosravi, S. Nahavandi, A survey of imitation learning: Algorithms, recent developments, and challenges, IEEE Transactions on Cybernetics (2024).
- [25] X. Huang, W. Liu, X. Chen, X. Wang, H. Wang, D. Lian, Y. Wang, R. Tang, E. Chen, Understanding the planning of llm agents: A survey, arXiv preprint arXiv:2402.02716 (2024).
- [26] Z. Huang, H. Liu, J. Wu, C. Lv, Differentiable integrated motion prediction and planning with learnable cost function for autonomous driving, IEEE transactions on neural networks and learning systems (2023).
- [27] Y. Zhu, S. Qiao, Y. Ou, S. Deng, N. Zhang, S. Lyu, Y. Shen, L. Liang, J. Gu, H. Chen, Knowagent: Knowledge-augmented planning for llm-based agents, arXiv preprint arXiv:2403.03101 (2024).
- [28] K. Yang, S. Li, M. Wang, X. Tang, Interactive decision-making integrating graph neural networks and model predictive control for autonomous driving, IEEE Transactions on Intelligent Transportation Systems (2025).
- [29] B. Ivanovic, A. Elhafsi, G. Rosman, A. Gaidon, M. Pavone, Mats: An interpretable trajectory forecasting representation for planning and control, arXiv preprint arXiv:2009.07517 (2020).
- [30] Z. Huang, H. Liu, C. Lv, Gameformer: Game-theoretic modeling and learning of transformer-based interactive prediction and planning for autonomous driving, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 3903–3913.
- [31] P. Karkus, B. Ivanovic, S. Mannor, M. Pavone, Diffstack: A differentiable and modular control stack for autonomous vehicles, in: Conference on robot learning, PMLR, 2023, pp. 2170–2180.
- [32] H. Liu, Z. Huang, W. Huang, H. Yang, X. Mo, C. Lv, Hybrid-prediction integrated planning for autonomous driving, IEEE Transactions on Pattern Analysis and Machine Intelligence (2025).
- [33] H. Liao, Y. Li, C. Wang, Y. Guan, K. Tam, C. Tian, L. Li, C. Xu, Z. Li, When, where, and what? a benchmark for accident anticipation and localization with large language models, in: Proceedings of the 32nd ACM International Conference on Multimedia, 2024, pp. 8–17.
- [34] S. Moon, H. Woo, H. Park, H. Jung, R. Mahjourian, H.-g. Chi, H. Lim, S. Kim, J. Kim, Visiontrap: Vision-augmented trajectory prediction guided by textual descriptions, arXiv preprint arXiv:2407.12345 (2024).
- [35] Y. Tang, H. Liao, T. Nie, J. He, A. Qu, K. Chen, W. Ma, Z. Li, L. Sun, C. Xu, E3ad: An emotion-aware vision-language-action model for human-centric end-to-end autonomous driving, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2026, pp. 10610–10620.
- [36] Y. Sun, N. Salami Pargoo, P. Jin, J. Ortiz, Optimizing autonomous driving for safety: A human-centric approach with llm-enhanced rlhf, in: Companion of the 2024 on ACM International Joint Conference on Pervasive and Ubiquitous Computing, 2024, pp. 76–80.
- [37] H. Liu, C. Li, Q. Wu, Y. J. Lee, Visual instruction tuning, Advances in neural information processing systems 36 (2023) 34892–34916.
- [38] H. Zhang, X. Li, L. Bing, Video-llama: An instruction-tuned audio-visual language model for video understanding, arXiv preprint arXiv:2306.02858 (2023).

- [39] I. Bae, J. Lee, H.-G. Jeon, Social reasoning-aware trajectory prediction via multimodal language model, *IEEE transactions on pattern analysis and machine intelligence* (2025).
- [40] X. Zheng, L. Wu, Z. Yan, Y. Tang, H. Zhao, C. Zhong, B. Chen, J. Gong, Large language models powered context-aware motion prediction, 2024. [arXiv:2403.11057](https://arxiv.org/abs/2403.11057).
- [41] L. Wen, D. Fu, X. Li, X. Cai, T. MA, P. Cai, M. Dou, B. Shi, L. He, Y. Qiao, Dilu: A knowledge-driven approach to autonomous driving with large language models, in: *The Twelfth International Conference on Learning Representations*, 2024. URL: <https://openreview.net/forum?id=0qTMUPuLuC>.
- [42] H. Liao, H. Shen, Z. Li, C. Wang, G. Li, Y. Bie, C. Xu, Gpt-4 enhanced multimodal grounding for autonomous driving: Leveraging cross-modal attention with large language models, *Communications in Transportation Research* 4 (2024) 100116.
- [43] S. Mandalika, A. Nambiar, et al., Primedrive-cot: A precognitive chain-of-thought framework for uncertainty-aware object interaction in driving scene scenario, in: *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 5293–5301.
- [44] H. Liao, Z. Li, H. Shen, W. Zeng, D. Liao, G. Li, C. Xu, Bat: Behavior-aware human-like trajectory prediction for autonomous driving, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 2024, pp. 10332–10340.
- [45] V. G. Kovvali, V. Alexiadis, L. Zhang, Video-based vehicle trajectory data collection, Technical Report, 2007.
- [46] R. Krajewski, J. Bock, L. Kloecker, L. Eckstein, The highd dataset: A drone dataset of naturalistic vehicle trajectories on german highways for validation of highly automated driving systems, in: *2018 21st international conference on intelligent transportation systems (ITSC)*, IEEE, 2018, pp. 2118–2125.
- [47] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, O. Beijbom, nuscenes: A multimodal dataset for autonomous driving, in: *2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, IEEE, 2020, pp. 11618–11628.
- [48] X. Huang, P. Wang, X. Cheng, D. Zhou, Q. Geng, R. Yang, The apolloscape open dataset for autonomous driving and its application, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42 (2020) 2702–2719.
- [49] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with bert, *arXiv preprint arXiv:1904.09675* (2019).
- [50] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, A. Alahi, Social gan: Socially acceptable trajectories with generative adversarial networks, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2255–2264.
- [51] N. Deo, M. M. Trivedi, Convolutional social pooling for vehicle trajectory prediction, in: *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 1468–1476.
- [52] T. Zhao, Y. Xu, M. Monfort, W. Choi, C. Baker, Y. Zhao, Y. Wang, Y. N. Wu, Multi-agent tensor fusion for contextual trajectory prediction, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12126–12134.
- [53] K. Gao, X. Li, B. Chen, L. Hu, J. Liu, R. Du, Y. Li, Dual transformer based prediction for lane change intentions and trajectories in mixed traffic environment, *IEEE Transactions on Intelligent Transportation Systems* (2023).
- [54] C. Tang, R. R. Salakhutdinov, Multiple futures prediction, *Advances in neural information processing systems* 32 (2019).
- [55] K. Messaoud, I. Yahiaoui, A. Verroust-Blondet, F. Nashashibi, Non-local social pooling for vehicle trajectory prediction, in: *2019 IEEE Intelligent Vehicles Symposium (IV)*, IEEE, 2019, pp. 975–980.
- [56] K. Messaoud, I. Yahiaoui, A. Verroust-Blondet, F. Nashashibi, Attention based vehicle trajectory prediction, *IEEE Transactions on Intelligent Vehicles* 6 (2021) 175–185.
- [57] X. Xie, C. Zhang, Y. Zhu, Y. N. Wu, S.-C. Zhu, Congestion-aware multi-agent trajectory prediction for collision avoidance, in: *2021 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 2021, pp. 13693–13700.
- [58] X. Chen, H. Zhang, F. Zhao, Y. Hu, C. Tan, J. Yang, Intention-aware vehicle trajectory prediction based on spatial-temporal dynamic attention network for internet of vehicles, *IEEE Transactions on Intelligent Transportation Systems* 23 (2022) 19471–19483.
- [59] X. Chen, H. Zhang, F. Zhao, Y. Cai, H. Wang, Q. Ye, Vehicle trajectory prediction based on intention-aware non-autoregressive transformer with multi-attention learning for internet of vehicles, *IEEE Transactions on Instrumentation and Measurement* 71 (2022) 1–12.
- [60] Z. Zuo, X. Wang, S. Guo, Z. Liu, Z. Li, Y. Wang, Trajectory prediction network of autonomous vehicles with fusion of historical interactive features, *IEEE Transactions on Intelligent Vehicles* (2023).
- [61] H. Liao, S. Liu, Y. Li, Z. Li, C. Wang, B. Wang, Y. Guan, C. Xu, Human observation-inspired trajectory prediction for autonomous driving in mixed-autonomy traffic environments, 2024. [arXiv:2402.04318](https://arxiv.org/abs/2402.04318).
- [62] P. Cong, Y. Xiao, X. Wan, M. Deng, J. Li, X. Zhang, Dacr-amp: Adaptive multi-modal vehicle trajectory prediction for dynamic drivable areas based on collision risk, *IEEE Transactions on Intelligent Vehicles* (2023).
- [63] H. Liao, Z. Li, G. Zhang, K. Li, C. Xu, Toward human-like trajectory prediction for autonomous driving: A behavior-centric approach, *Transportation Science* 59 (2025) 823–852.
- [64] C. Wang, H. Liao, K. Zhu, G. Zhang, Z. Li, A dynamics-enhanced learning model for multi-horizon trajectory prediction in autonomous vehicles, *Information Fusion* 118 (2025) 102924.
- [65] Y. Cai, Z. Wang, H. Wang, L. Chen, Y. Li, M. A. Sotelo, Z. Li, Environment-attention network for vehicle trajectory prediction, *IEEE Transactions on Vehicular Technology* (2021).
- [66] H. Cui, V. Radosavljevic, F.-C. Chou, T.-H. Lin, T. Nguyen, T.-K. Huang, J. Schneider, N. Djuric, Multimodal trajectory predictions for autonomous driving using deep convolutional networks, 2019. [arXiv:1809.10732](https://arxiv.org/abs/1809.10732).
- [67] Y. Chai, B. Sapp, M. Bansal, D. Anguelov, Multipath: Multiple probabilistic anchor trajectory hypotheses for behavior prediction, 2019. [arXiv:1910.05449](https://arxiv.org/abs/1910.05449).
- [68] T. Phan-Minh, E. C. Grigore, F. A. Boulton, O. Beijbom, E. M. Wolff, Covernet: Multimodal behavior prediction using trajectory sets, in: *Proceedings of the CVPR*, 2020, pp. 14074–14083.
- [69] T. Salzmann, B. Ivanovic, P. Chakravarty, M. Pavone, Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data, in: *ECCV*, Springer, 2020, pp. 683–700.

- [70] Y. Yuan, X. Weng, Y. Ou, K. Kitani, Agentformer: Agent-aware transformers for socio-temporal multi-agent forecasting, 2021. [arXiv:2103.14023](#).
- [71] B. Kim, S. H. Park, S. Lee, E. Khoshimjonov, D. Kum, J. Kim, J. S. Kim, J. W. Choi, Lapred: Lane-aware prediction of multi-modal future trajectories of dynamic agents, 2021. [arXiv:2104.00249](#).
- [72] Z. Zhong, Y. Luo, W. Liang, Stgm: Vehicle trajectory prediction based on generative model for spatial-temporal features, *IEEE TITS* 23 (2022) 18785–18793.
- [73] P. Xu, J.-B. Hayet, I. Karamouzas, Context-aware timewise vaes for real-time vehicle trajectory prediction, *IEEE Robotics and Automation Letters* (2023).
- [74] Y. Xu, Y. Fu, Adapting to length shift: Flexilength network for trajectory prediction, in: *Proceedings of the CVPR*, 2024, pp. 15226–15237.
- [75] C. Wong, B. Xia, Z. Zou, Y. Wang, X. You, Socialcircle: Learning the angle-based social interaction representation for pedestrian trajectory prediction, in: *Proceedings of the CVPR*, 2024, pp. 19005–19015.
- [76] Q. Zhang, Y. Yang, P. Li, O. Andersson, P. Jensfelt, Seflow: A self-supervised scene flow method in autonomous driving, *arXiv preprint arXiv:2407.01702* (2024).
- [77] B. Rao, H. Liao, Y. Guan, C. Wang, B. Wang, J. Zhang, Z. Li, Amd: Adaptive momentum and decoupled contrastive learning framework for robust long-tail trajectory prediction, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 28849–28858.
- [78] C. Wang, H. Liao, Z. Li, C. Xu, Wake: Towards robust and physically feasible trajectory prediction for autonomous vehicles with wavelet and kinematics synergy, *IEEE TPAMI* (2025).
- [79] R. Girgis, F. Golemo, F. Codevilla, M. Weiss, J. A. D’Souza, S. E. Kahou, F. Heide, C. Pal, Latent variable sequential set transformers for joint multi-agent motion prediction, in: *International Conference on Learning Representations*, 2022. URL: [https://openreview.net/forum?id=Dup\\_dDqkZC5](https://openreview.net/forum?id=Dup_dDqkZC5).
- [80] O. Scheel, L. Bergamini, M. Wolczyk, B. Osiński, P. Ondruska, Urban driver: Learning to drive from real-world demonstrations using policy gradients, in: *CoRL*, 2022, pp. 718–728.
- [81] B. Hui, J. Yang, Z. Cui, J. Yang, D. Liu, L. Zhang, T. Liu, J. Zhang, B. Yu, K. Lu, et al., Qwen2. 5-coder technical report, *arXiv preprint arXiv:2409.12186* (2024).
- [82] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma, et al., Scaling instruction-finetuned language models, *Journal of Machine Learning Research* 25 (2024) 1–53.
- [83] P. Zhang, G. Zeng, T. Wang, W. Lu, Tinyllama: An open-source small language model, 2024. [arXiv:2401.02385](#).
- [84] Y. Li, S. Bubeck, R. Eldan, A. D. Giorno, S. Gunasekar, Y. T. Lee, Textbooks are all you need ii: phi-1.5 technical report, 2023. [arXiv:2309.05463](#).